Genome-wide association study of copy number variation in 16,000 cases of eight common diseases and 3,000 shared controls

Supplementary Information

Version 3.0.0

The Wellcome Trust Case Control Consortium*

March 4, 2010

Orientation to Supplementary Material

This supplementary material is designed to be read in conjunction with the main published text. It provides additional relevant background and experimental information with a level of detail that cannot be provided within the main text for reasons of space constraints. It covers details of the pilot study that informed the main experiment, background information about the phenotypes and samples, and details of the methodology, quality control, data properties and results. We have also provided a glossary of the commonly used abbreviations and acronyms that appear within the paper. Where necessary additional information can be obtained from the corresponding author or relevant disease investigators.

Contents

*List of participants and affiliations appear at the end of the Main paper.

Figures and tables referenced from main paper The following supplementary figures are referenced in the main paper: 9, 18, 19, 20, 21, 22, 23, 27, 29, 30

The following supplementary tables are referenced in the main paper: 13, 14

The following supplementary figure is referenced in the online methods: 24

The following supplementary tables are referenced in the online methods: 2, 7, 10, 15

1 Pilot study

1.1 Design

To identify the best CNV genotyping array for this study we conducted a pilot study. This pilot study genotyped four 96 well plates (30 HapMap CEPH trios, UKBS, RA and T1D, for a total of 378 samples) that were typed on a set of 156 previously identified CNVs. These collections were selected to span the range of DNA qualities observed in the WTCCC1 SNP genotyping study, assessed by the SNP call rate from that study. These CNVs came from five different sources:.

- a survey of deletions in the human genome based on HapMap trio data[1].
- a genome-wide scan of 270 HapMap individuals[2] using a CGH assay developed at the Sanger Institute[3].
- a second genome wide scan of the same 270 HapMap individuals using the Affymetrix 500K early access array[2].
- a set of CNVs identified by analysing the data from the initial WTCCC1 Affymetrix 500K study.
- three manually selected multi-allelic CNVs previously reported in the literature.

Three genotyping platforms were compared:

- The 7k Illumina iSelect format, an average of 40 probes per CNV and no probe replicates.
- The 105k Agilent CGH format: 60 probes on average per CNV and 10 replicates for each probe.
- The 135k NimbleGen CGH format, consisting of 141,001 probes: 90 distinct probes for each CNV on average and 10 replicates for each probe.

Due to technical constraints in array design the number of probes on the Illumina iSelect array was significantly lower than Agilent/Nimblegen for the pilot study. However, this difference is not relevant to the final array design as for a large scale order of the size of the WTCCC study the design cost becomes less significant. All three proposed final formats used a comparable numbers of probes, approximately equal to 100,000.

The 156 CNVs targeted on the arrays were mostly drawn from CNV discovery projects in the four HapMap populations and may not have been polymorphic in our pilot samples. We identified polymorphic CNVs by measuring correlations in signal across platforms, on the basis that such correlations would result from differences in CNV genotypes. The threshold for the correlation coefficient was set to 0.2. This value was obtained by computing the distribution of correlation coefficients for 67 of the 156 CNVs that were known to be polymorphic (based on a previous study[4]). From this analysis we identified 108 polymorphic CNVs out of 156 CNVs on the array.

1.2 Data analysis

To analyze the data from this pilot experiment, we defined a binary outcome variable for each CNV. A CNV was labelled as successfully genotyped when the loss of effective sample size caused by ambiguous genotyping did not exceed 10%. These estimates of the effective statistical power rely on evaluating the asymptotic variance of the estimated odds ratio (see[5] for details). This outcome variable is a function of the number of probes and we always selected the best available probes as defined by the manufacturer.

To allow data comparison across multiple samples, we first normalised the data. For both the Agilent and the Nimblegen CGH arrays, we computed the log ratio of the sample DNA (green) to the control DNA (red) and the distribution of log-ratios were then quantile normalised against a unique reference distribution. We considered alternative normalization strategies, including separate normalizations of the green/red channel, using the green channel only, or simply using the log2 ratio without any additional normalization step. We found that different normalizations led to better or worse signals on a cnv-by-cnv basis but that no one normalization scheme performed universally better. For the Illumina genotyping arrays, we used for each sample median inter-quartile range normalisation to set the median intensity signal to 0 and the 25%-75% range to 1. Following this normalisation step, several methods were considered to summarise intensity data across multiple probes. We investigated the use of principal component, combined with a second step linear discriminant analysis (see[5] for details) and at many (but not all CNVs) this improved the signal. Last, we manually refined our estimates of CNV boundaries in order to exclude CNV probes not located in the copy number variable DNA region.

Note that the Illumina array provided data for two types of probes: CNV and SNP probes. In contrast with CNV probes, SNP probes also provide genotype call data. For example, loss of SNP heterozygosity is

expected in the presence of a deletion. To facilitate comparison across platforms, and because of the analytical challenges of combining CNV and SNP data, our analysis of the Illumina pilot data only used the overall sum intensity at all CNV and SNP probes but did not take advantage of the SNP genotype information.

1.3 Results

We summarized the genotyping accuracy using the fraction of successfully genotyped CNVs. We found the data quality to be highly variable across CNVs and genotyping arrays. Bivariate plots shown in Supplementary Figure 1 give examples of the type of CNV data that were analysed.

The number of clusterable CNVs is a function of two parameters: the number of distinct probes per CNV, and the number of replicates for each of these probes. While the success rate increased with the number of probes per CNV, this proportion reached a maximum for all three arrays at about 10 probes. We obtained the highest success rate by maximising the number of distinct probes rather than replicating a smaller subset of better quality probes (see Supplementary Figure 2). Probe quality metrics defined independently by the three suppliers for their own platforms proved to be useful in ensuring optimal data quality for each CNV locus (see Supplementary Figure 3). The highest genotyping rate was obtained for deletions and the worst for multi-allelic CNVs.

The CNV genotyping success rate was, however, nearly independent of the CNV size (see Supplementary Table 1 and Supplementary Figure 4).

Across the full range of parameters considered the Agilent array provided the highest success rate. Using the Agilent array, 10 distinct probes per CNV and no replicate probe we were able to genotype 68% of deletions, 64% of duplications and 44% of multi-allelic CNVs among the 108 polymorphic CNVs in the pilot study (see Supplementary Table 1).

For the two CGH platforms, while over all CNVs we observed poorer clustering on the green channel intensity (test sample) as compared to clustering of the ratio of (green/red), we identified three CNVs where clustering on the green channel intensity was markedly improved over clustering on the log2 ratio of (green/red). Further analysis of these three loci revealed that at all three of these loci the reference sample is homozygously deleted at this locus, and thus the red channel intensity represents background signal only. This motivates the use of a pooled reference sample in CGH-based genotyping experiments to minimize the number of loci at which all reference chromosomes have the deleted allele.

2 Samples

The WTCCC CNV study analysed cases from 8 common diseases (Breast Cancer (BC), Bipolar Disorder (BD), Coronary Artery Disease (CAD), Crohn's Disease (CD), Hypertension (HT), Rheumatoid Arthritis (RA), Type I Diabetes (T1D), and Type 2 Diabetes (T2D)) and two control cohorts (1958 Birth Cohort (58C) and the UK Blood Service collection (UKBS)). Except where described explicitly below, neither cases nor controls were karyotyped. The number of subjects from each cohort that were analysed and the numbers that passed each phase of the quality control (QC) procedures within this study are shown in Supplementary Table 8. For BD, CAD, CD, HT, RA, T1D, T2D, and the two control cohorts, a large proportion of the subjects studied in this experiment were the same as those in the WTCCC1 SNP genome-wide association study (GWAS) (Supplementary Table 2). Where sufficient DNA was not available for the original WTCCC1 individuals, additional new samples from the same cohorts were used, selected using the same approaches used for the WTCCC1 samples. Any samples that failed any of the relevant QC metrics in WTCCC1 were excluded from consideration for this experiment. The BC cohort was not included in the WTCCC1 SNP GWAS study. Brief details about each disease and the current knowledge about its genetic architecture (including any information about CNVs) together with information about the recruitment and ascertainment strategy for cohort are provided below. Appropriate local and/ or national ethical approval is held by the principal investigators (PIs) of each cohort.

2.1 Breast cancer (BC)

BC Phenotype and genetic findings to date - Breast Cancer is a common disease affecting approximately one in ten women in developed countries. Overall BC is twice as common in women with an affected first degree relative[6,7]. Twin studies demonstrate a substantially higher risk to monozygotic twins of affected relatives than to dizygotic twins, suggesting that the familial aggregation is predominantly due to genetic factors rather than shared environmental factors; the markedly skewed distribution of genetic liability in twins suggests that the majority of genetic risk may lie within a genetically predisposed minority[8]. Genome-wide linkage analysis and positional cloning have identified two high penetrance breast cancer predisposition genes, *BRCA1* and *BRCA2*, mutations in which are rare, cause loss-of-function and confer a high risk of breast cancer (Relative risk, RR$\geq$10)[6,9,10]. Mutational screening of genes functionally related to *BRCA1* and/or *BRCA2* has revealed four genes, *CHEK2*, *ATM*, *BRIP1* and *PALB2*; mutations in these genes are also rare and cause loss-of-function but confer a more modest risk of breast cancer (RR 2-4)[11,12,13,14]. Association studies have identified, at genome-wide levels of statistical significance, a further thirteen common variants associated with breast cancer predisposition (per allele RR$\leq$1.3); twelve of these were identified through genome-wide association studies[15,16,17,18,19,20,21,22]. The diverse array of rare mutations detected in *BRCA1* and *BRCA2* includes a number of exonic deletions and duplications; to date there has been no robust evidence to implicate common copy number variants in breast cancer predisposition.

BC sample description - Breast Cancer samples were from independently ascertained women with invasive breast cancer, each of whom had a family history of breast cancer in relatives. The samples were from breast cancer families recruited through Cancer Genetics clinics in the UK; families from non-UK ethnic groups were excluded. We quantified the extent of the family history of breast cancer using a Family History Score, which was defined as the number of relatives of the index case with breast cancer, weighted by their degree of relatedness to the index case to adjust for the expected allele sharing (score=1 for a proband affected with unilateral breast cancer, score=2 for a proband affected with bilateral breast cancer; 0.5 was added for each affected 1st degree relatives and 0.25 for 2nd degree relatives; relatives with bilateral breast cancers score double). The range of family history scores was 1.5-5.25; over 98% of samples had a score of $\geq$ 1.75 (i.e. a woman with breast cancer with one first- and one second-degree relatives or equivalent). We excluded mutations in the full coding sequence and intron/exon boundaries of *BRCA1* and *BRCA2* in over 95% of samples via Conformation Sensitive Gel Electrophoresis or direct sequencing. We also performed Multiplex ligation-dependent Probe Amplification (MLPA) analysis on these samples using the SALSA MLPA KIT P002B *BRCA1* kit and SALSA MLPA KIT P045-B1 *BRCA2/CHEK2* kit (MRC Holland) and the manufacturers' protocols. We obtained informed consent from all patients and the research was approved by the London Multicentre Research Ethics Committee (MREC/01/2/18).

2.2 Bipolar disorder (BD)

BD Phenotype and genetic findings to date - Bipolar disorder (BD; manic depressive illness[23]) refers to an

episodic recurrent pathological disturbance in mood (affect) ranging from extreme elation or mania to severe depression and usually accompanied by disturbances in thinking and behaviour: psychotic features (delusions and hallucinations) often occur. Pathogenesis is poorly understood, and the phenotype is based solely on clinical features. There is robust evidence for a substantial genetic contribution to risk of BD[24] and for an overlap in genetic susceptibility of BD and schizophrenia[25]. The estimated sibling recurrence risk (λ_s) is 7-10 and heritability 80-90%[26,24]. To date 2 loci have been reported at genome-wide levels of statistical significance in meta-analyses of GWAS samples[27] and there is evidence for an overlap in genetic susceptibility between BD and schizophrenia[28,29,30]. The aggregate burden ("load") of large, rare Copy Number Variants is increased in schizophrenia cases compared with controls[31,32,33] and some specific rare, large CNV loci have been implicated in both schizophrenia and autism[34]. In contrast, no general increased burden of large, rare CNVs has yet been observed in BD compared with controls[35,36] and the burden of very large (>1Mb), rare CNVs has been reported to be significantly lower in bipolar disorder than schizophrenia[36].

BD sample description - BD cases were all over the age of 16 years, living in mainland UK and of European descent. Recruitment was undertaken throughout the UK by teams based in Aberdeen, Birmingham, Cardiff, London and Newcastle. Individuals who had been in contact with mental health services were recruited if they suffered with a major mood disorder in which clinically significant episodes of elevated mood had occurred. This was defined as a lifetime diagnosis of a bipolar mood disorder according to Research Diagnostic Criteria[37] and includes manic disorder, bipolar I disorder, bipolar II disorder and schizoaffective disorder bipolar type. After providing written informed consent, all subjects were interviewed by a trained psychologist or psychiatrist using a semi-structured lifetime diagnostic psychiatric interview (in most cases the Schedules for Clinical Assessment in Neuropsychiatry[38]) and available psychiatric medical records were reviewed. Using all available data, best-estimate ratings were made for a set of key phenotypic measures which included as a minimum the OPCRIT checklist (which covers 90 items of psychopathology and course of illness)[39,40] and lifetime psychiatric diagnoses were assigned according to the Research Diagnostic Criteria[37]. The reliability of these methods has been shown to be high[41,42]. Further details of clinical methodology can be found in[41,42].

2.3 Coronary artery disease (CAD)

CAD Phenotype and genetic findings to date - Coronary artery disease (coronary atherosclerosis) is a chronic degenerative condition in which lipid and fibrous matrix is deposited in the walls of the coronary arteries to form atheromatous plaques[43]. It may be clinically silent or present with angina pectoris or acute myocardial infarction. Pathogenesis is complex, with endothelial dysfunction, oxidative stress and inflammation contributing to development and instability of the atherosclerotic plaque[43]. In addition to lifestyle and environmental factors, genes are important in the aetiology of CAD[44]. For early myocardial infarction, estimates of λ_s in the range from $\sim$2 to $\sim$7[45]. Genetic variation is thought likely to influence risk of CAD both directly and through effects on known CAD risk factors including hypertension, diabetes and hypercholesterolaemia.

Genome-wide association analyses have, so far, identified 12 loci associated with risk of CAD which have been validated in additional samples[46,47,48,49,50,51,52]. None of these loci harbour known CNVs. Using data extracted from the Affymetrix 6.0 GeneChip, the MIGen Consortium have recently also tested the association of 554 common copy number polymorphisms (>1% allele frequency) with risk of premature MI in approximately 3000 cases and 3000 controls. None of the CNVs met their pre-specified threshold for replication of $P < 10^{-3}$. They also identified 8,065 rare CNVs but did not detect a greater CNV burden in cases compared to controls.[50].

CAD sample description - CAD cases had a validated history of either myocardial infarction or coronary revascularisation (coronary artery bypass surgery or percutaneous coronary angioplasty) before their 66th birthday. Verification of the history of CAD was required either from hospital records or the primary care physician. Recruitment was carried out on a national basis in the UK through (i) a direct approach to the public via the media and (ii) mailing all general practices (family physicians) with information about the study, as previously described[53]. In an initial pilot phase, potential participants were also identified and approached through local CAD databases in the two lead centres (Leeds and Leicester). Whilst the majority of subjects had at least one further sib also affected with premature CAD, only one subject from each family was included in the present study.

2.4 Crohn's disease (CD)

CD Phenotype and genetic findings to date - Crohn's disease is a common form of chronic inflammatory

bowel disease[54]. The pathogenic mechanisms are not well understood, but a genetic contribution is suggested by a λ_s of 17-35 and by twin studies that contrast monozygotic concordance rates of 50% with only 10% in dizygotic pairs[55,56]. This has been confirmed by a series of genome-wide association studies (GWAS) in CD (reviewed in[57]) and a meta-analysis which indicates the involvement of at least 30 distinct susceptibility loci[58]. The GWAS have also provided important insights into pathogenesis, highlighting the regulation of the interleukin-23 and Th17 cell pathway and of bacterial clearance by autophagy as potentially important components[59,60]. No systematic association study of copy number variation in CD has yet been undertaken, but there is an unconfirmed report of a reduction in copy number of the Beta-Defensin 2 gene (HBD-2) in colonic CD[61]. More recently, a CNV just upstream of the IRGM gene has been shown to be highly correlated with SNPs at this locus that are strongly associated with CD and may influence IRGM expression[62].

CD sample description - CD patients were unrelated, white, European attendees at inflammatory bowel disease clinics in and around the five centres which contributed samples to the WTCCC (Cambridge, Oxford, London, Newcastle, Edinburgh). Ascertainment was based on a confirmed diagnosis of Crohn's disease (CD) using conventional endoscopic, radiological and histopathological criteria[63]. We included all sub-types of CD as classified by disease extent and behaviour and the cohort was not specifically enriched for family history or early age of onset. All patients provided written consent and a sample of blood, from which DNA was extracted by standard protocols.

2.5 Hypertension (HT)

HT Phenotype and genetic findings to date - Hypertension refers to a clinically significant increase in blood pressure and constitutes an important risk factor for cardiovascular disease (`http://www.who.int/whr/2002/en/`;[64]). Lifestyle exposures that elevate blood pressure, including sodium intake, alcohol and excess weight[65] are well-described risk factors. Genetic factors are also important[66], and to date there are 13 loci influencing systolic, diastolic blood pressure and hypertension risk which have been robustly validated. These loci were detected by meta-analyses of genomewide association scans, and are common variants that exhibit modest effects on the phenotype[67,68]. There are also reports of rare variants within genes identified by the study of Mendelian forms of hypertension influencing blood pressure in the general population[69,70]. To date there are no reports of copy number variants associated with hypertension or blood pressure phenotypes. The estimates for hypertension of λ_s are approximately 2.5-3.5.

HT sample description - HT cases comprised severely hypertensive probands ascertained from families with multiplex affected sibships or as parent-offspring trios. They were of white British ancestry (up to level of grandparents) and were recruited from the Medical Research Council General Practice Framework and other primary care practices in the UK[71]. Each case had a history of hypertension diagnosed prior to 60 years of age, with confirmed blood pressure recordings corresponding to seated levels $>$150/100 mmHg (if based on one reading), or the mean of 3 readings greater than 145/95 mmHg. These criteria correspond to the threshold for the uppermost 5% of blood pressure distribution in a contemporaneous health screening survey of 5000 British men and women in 1995 (N. Wald and M. Law, personal communication). We excluded hypertensive individuals who self-reportedly consumed $>$21 units of alcohol/week and those with diabetes, intrinsic renal disease, a history of secondary hypertension or co-existing illness. We focused on the recruitment of hypertensive individuals with BMI $<$30kgm^{-2}. The probands were extensively phenotyped by trained nurses (see `http://www.brightstudy.ac.uk` for standard operating procedures, additional phenotypes and study questionnaires). Sample selection for WTCCC was based on DNA availability and quality.

2.6 Rheumatoid arthritis (RA)

RA Phenotype and genetic findings to date - Rheumatoid arthritis is a chronic inflammatory disease characterized by destruction of the synovial joints resulting in severe disability, particularly in patients who remain refractory to available therapies[72]. Susceptibility to, and severity of, RA are determined by both genetic and environmental factors, with λ_s estimates ranging from 5-10[73].

The advent of genome-wide association technologies has led to rapid advances in the identification of a number of RA susceptibility loci. The two major susceptibility loci are the HLA DRB1 and PTPN22 genes but at least 14 other loci have been confidently confirmed in multiple populations (reviewed in ref. 74). The majority of the associated SNP variants map within or close to genes with immunological functions and many of the loci are also associated with other autoimmune diseases. In total, the

SNP variants identified to date are estimated to account for less than half of the total genetic contribution to RA.

Of interest for RA are the recent claims of association to autoimmune and inflammatory-mediated disease of three CNV loci located at chromosomes 17q11-q12, 8p23 and 1q23. The 17q11-q12 locus contains a 90 Kb segmental duplication containing two transcripts, CC chemokine ligand 3 like 1 (*CCL3L1*) and *CCL4L1*, both potent ligands for CC chemokine receptor 5 (*CCR5*). This region is of particular significance as increased copy number of the duplication has been associated with susceptibility to RA[75]. A 250 kb variable repeat mapping to chromosome 8p23 has been associated with Crohn's disease and psoriasis[61,76]. This region contains multiple genes from the beta defensin gene family, known for having antibacterial, antiviral and chemokine-like activity. The identification of Fc fragment of IgG low affinity IIIb receptor (*FCGR3B*), located on chromosome 1q23, as a candidate CNV for autoimmune disease originated from work conducted in experimental rat models of glomerulonephritis[77]. A further study has demonstrated that variation in the region is associated with systemic autoimmune disorders, such as SLE, microscopic polyangiitis and Wegener's granulomatosis[78].

RA sample description - RA cases were recruited to studies co-ordinated by the **arc** Epidemiology Unit. All subjects were Caucasian over the age of 18 and satisfied the 1987 American College of Rheumatology Criteria for RA[79] modified for genetic studies[80].

2.7 Type 1 diabetes (T1D)

T1D Phenotype and genetic findings to date - Type 1 diabetes is a chronic autoimmune disorder with onset usually in childhood[81]. Over 40 genetic loci are convincingly associated with T1D[82,83,84,46,85] and twin data suggest that over 85% of the phenotypic variance is due to genetic factors[86]. The λ_s for T1D has been estimated to be around 15, although more recent analyses suggest this in an exaggeration, and that λ_s may be less than ten with the currently known loci explaining a λ_s of just under five in the UK population[87]. To date, no evidence of association of CNVs with T1D has been reported[88].

WTCCC T1D sample description - T1D cases were recruited from paediatric and adult diabetes clinics at 150 National Health Service hospitals across mainland UK. The total T1D case dataset (N=8000) from which the WTCCC cases were selected, represents close to half the T1D cases seen in such clinics. Nationwide coverage was achieved through the voluntary efforts of members of the British Society for Paediatric Endocrinology and Diabetes, who recruited about half of cases, the rest coming from peripatetic nurses employed by the JDRF/WT GRID project[89]. To establish a positive diagnosis of T1D (and, in particular, to distinguish it from the more common, but later onset T2D), we required all cases to have an age of diagnosis below 17 and insulin dependence since diagnosis (with a minimum period of at least 6 months)[90]. However, a few subjects were subsequently discovered to be suffering from rare monogenic disorders, such as maturity onset diabetes of the young (MODY)[91], and latterly permanent neonatal diabetes (PNDM)[92]: these were excluded.

2.8 Type 2 diabetes (T2D)

T2D Phenotype and genetic findings to date - Type 2 diabetes is a chronic metabolic disorder which is generally first diagnosed in the middle to late adulthood[93]. Strongly associated with obesity, individuals with established T2D display defects in both the secretion and peripheral actions of insulin[94]. The familial aggregation of T2D (an estimated λ_s of $\sim$3.0 in European individuals)[65] reflects both shared family environment and genetic predisposition. Heritability values vary widely with most estimates between 30 and 70%[94]. Largely as a result of genome-wide association efforts, the number of confirmed T2D-susceptibility loci stands at around 20[95]. The effect sizes associated with these common risk variants are modest (the largest allelic odds ratio is $\sim$1.35) such that known risk variants capture less than 10% of observed familiality[95,96]. There is no evidence to date to implicate common copy number polymorphisms in T2D risk, though (a) a common deletion near NEGR1 has recently been implicated in variation in adult body mass index[97] and (b) rare deletions of the HNF1B gene cause some forms of maturity onset diabetes of the young (MODY)[98].

T2D sample description - The T2D cases were selected from UK Caucasian subjects who form part of the Diabetes UK Warren 2 repository. In each case, the diagnosis of diabetes was based on either current prescribed treatment with sulphonylureas, biguanides, other oral agents and/or insulin or, in the case of individuals treated with diet alone, historical or contemporary laboratory evidence of hyperglycemia (as defined by World Health Organization). Other forms of diabetes (e.g., maturity-onset diabetes of the young, mitochondrial diabetes, and type 1 diabetes) were excluded by standard clinical criteria based on personal and family history.

Criteria for excluding autoimmune diabetes included absence of first-degree relatives with T1D, an interval of $\geq$1 year between diagnosis and institution of regular insulin therapy and negative testing for antibodies to glutamic acid decarboxylase (anti-GAD). Cases were limited to those who reported that all four grandparents had exclusively British and/or Irish origin, by both self-reported ethnicity and place of birth. All were diagnosed between age 25 and 75. Approximately 25% were explicitly recruited as part of multiplex sibships[99] and $\sim$20% were offspring in parent-offspring "trios" or "duos" (that is families comprising only one parent complemented by additional sibs)[100]. The remainder were recruited as isolated cases but these cases were (compared to population-based cases) or relatively early onset and had a high proportion of T2D parents and/or siblings[101]. Cases were ascertained across the UK but were centred around the main collection centres (Exeter, London, Newcastle, Norwich, Oxford). Selection of the samples typed in WTCCC from the larger collections was based primarily on DNA availability and success in passing WTCCC DNA QC.

2.9 1958 Birth Cohort (58C)

The common control groups for the WTCCC were derived from two sources. The first was made up of 1500 individuals from the British 1958 Birth Cohort (also known as the National Child Development Study) which includes all births in England, Wales and Scotland during one week in 1958, (`http://www.cls.ioe.ac.uk/studies.asp?section=000100020003`)[102]. Survivors have been followed up by parental interview and school medical examination at ages 7, 11 and 16, and by cohort member interview at 23, 33 and 41 years. Immigrants of the same dates of birth were identified at ages 7, 11 and 16, and followed into adulthood, but adult immigrants (after age 16) have not been included. DNA was extracted from cell lines which had been previously grown as part of the resource.

2.10 UK Blood Services Controls (UKBS)

The second set of common controls was made up of 1500 individuals selected from a sample of blood donors recruited as part of the current project. WTCCC in collaboration with the UK Blood Services (NHSBT in England, SNBTS in Scotland and WBS in Wales) set up a UK national repository of anonymised samples of DNA and viable mononuclear cells from 3622 consenting blood donors, age range 18–69 years (ethical approval 05/Q0106/74). A set of 1564 samples was selected from the 3622 samples recruited based on sex and geographical region (to reproduce the distribution of the samples of the 1958 Birth Cohort) for use as common controls in the WTCCC study.

3 CNV experiment

3.1 Design of CNV genotyping array

The primary objective of the design of the CNV-typing array was to maximise the number of common CNVs assayed on the array by combining information from multiple sources of discovered CNVs in a complementary fashion. Secondary objectives were to: (i) screen for rare CNVs in exons of a limited number of genes selected on a disease-by-disease basis, and (ii) assess CNV at loci highlighted through CNV analysis of the Affymetrix 500k data generated by the WTCCC1 experiment on seven of the same diseases. An overview of the array contents is given in Supplementary Figure 5.

Design parameters

Analysis of the data from the pilot study allowed us to define key parameters of the array design:

1. That among the technology platforms that were assessed, the Agilent CGH platform allowed the greatest proportion of CNVs to be genotyped.

2. That on the Agilent CGH platform there is little additional genotyping power to be gained from placing more than 10 probes in each CNV.

3. That Agilent's *in silico* probe performance score is a meaningful metric for prioritising probes for inclusion on the array

4. That replicating the probes with the best *in silico* performance scores does not increase power to genotype CNVs, but neither does a limited amount of probe replication lessen power.

The ∼105,000 probe Agilent CGH array format was chosen on the basis that compared to the smaller (∼44,000 probe) or larger (∼240,000 probe) array formats it represented the most cost-efficient option, allowing inclusion of almost all targeted loci on the array.

Targeted loci

The loci targeted on the array fell into four main categories:

1. Control loci – to facilitate identification of sample mishandling:

   (a) X-chromosomal probes allowing sex to be determined [0.09% of targeted loci]

   (b) CNV loci genotyped in the WTCCC1 Affymetrix 500k data data [0.15% of targeted loci]

2. CNV loci – loci thought to vary in copy number:

   (a) Loci from the microarray-based CNV discovery project conducted by the Genome Structural Variation (GSV) Consortium[103] [83.62% of targeted loci].

   (b) Polymorphic CNVs not present in the GSV map above, identified from CNV analyses of SNP genotyping chips

      i. From analyses of Affymetrix 6.0 chip data in the HapMap CEU[104] [0.71% of targeted loci]

      ii. From analyses of Illumina 1M chip data in the HapMap populations (to be described in a forthcoming manuscript) [1.96% of targeted loci]

      iii. From analyses of the Affymetrix 500k data in the WTCCC1 experiment (to be described in a forthcoming manuscript)

   (c) Insertions of novel sequence not present in the reference sequence [2.51% of targeted loci]

      i. Loci from assembly of unmapped fosmid end sequences[105]

      ii. Loci from comparison of the Venter genome assembly and the reference genome assembly[106]

3. Candidate genes - candidate exons for rare, potentially high penetrance alleles for each disease [7.90% of targeted loci]

4. Validation loci - loci exhibiting signals of CNV association from analyses of the WTCCC1 Affymetrix 500k data (to be described in a forthcoming manuscript) [2.36% of targeted loci]

These different classes of loci are described in more detail below, and the loci targeted on the final array are listed in Supplementary Table 3.

Note that the major source of loci targeted on the array are those from the GSV microarray-based CNV discovery project. A preliminary set of loci shared by the GSV consortium was used to design the array. In the situation where loci overlap one another, the longest contiguous subregion most unique to each locus was identified for probe design. Supplementary Table 3 reports the

most unique subregions for each of this preliminary set of GSV loci. These loci definitions were subsequently refined by the GSV consortium, and the probes on the array mapped onto these new loci definitions for all the analyses reported in this study. Thus the loci reported in Supplementary Table 3 do not precisely correspond to the loci reported elsewhere in this study, and by the GSV consortium.

Probe Design

To maximise the proportion of the 105,072 probes allocated to CNV content on the array the number of Agilent's standard control probes was cut from 4,626 to 1,060, by including only the orientation controls, required for image orientation and negative controls required for background subtraction by Agilent's Feature Extraction software.

Probes within targeted loci were first sought within the Agilent's existing database of pre-designed probes (then totalling 24 million probes). The target number of different probes could be identified for 65% of all targeted loci, and 89% of targeted loci had at least one probe. One round of custom probe design was then undertaken for the remaining 35% of loci for which insufficient numbers of probes were available from the existing database, after which 84% of targeted loci had sufficient numbers of different probes. A second round of custom probe design was performed using a less stringent homology filter (WindowsMasker, rather than RepeatMasker) to maximise the number of different probes that could be designed to the remaining 16% of targeted loci.

For loci where CNV genotyping is the primary goal (i.e. CNV loci and validation loci), 10 different probes within the locus were sought. If more than 10 could be identified, the ten with the top in silico probe performance score were selected. If fewer than 10 probes but more than 4 probes could be designed then the top-scoring probes were duplicated on the array to bring the total number of probes up to 10. If 4 or fewer probes could be designed then the locus was discarded.

For loci where CNV discovery is the primary goal (i.e. candidate exons), 3 probes passing Agilent's standard QC metrics were sought for each exon. If more than 3 probes could be designed then the 3 top scoring probes were selected. If only 2 probes could be designed then the top-scoring probe was duplicated, if only 1 probe could be designed then it was triplicated.

As a result of this hierarchical design procedure, it was possible to include the target number of probes in 94% of the targeted loci. The breakdown by source of target loci is shown in the Supplementary Table 3. The sources of targeted loci, and the numbers of probes on the array are summarised in Figure 5.

Detailed description of targeted loci

Control loci We considered two classes of control loci to facilitate detection of sample mishandling:

1. X-chromosomal probes in CNV deserts to allow individuals to be sexed on the basis of X/autosomal intensity ratios. We identified 10 X-chromosomal CNV deserts in the CNV map from WTCCC1 Affymetric 500k CNV analyses and placed 3 probes in each.

2. Common CNVs that are genotypable from the WTCCC1 Affymetric 500k data, for which CNV genotypes could be checked against those from WTCCC1 data. We identified 18 CNVs that could be reasonably well-genotyped from the WTCCC1 data (good cluster separation) for inclusion on the WTCCC array design, with 10 probes in each CNV.

CNV loci: GSV discovery project The GSV CNV discovery project is described fully elsewhere[103], and is summarised below. A set of 20 arrays comprising 42,000,000 probes tiling across the assayable portion of the human reference genome (median probe spacing 56bp), was used in array-CGH experiments on 40 individuals (including 20 HapMap CEU samples of European ancestry and 20 HapMap YRI samples of West African ancestry) against a common reference sample. CNV calls were made individual-by-individual on all 40 experiments, requiring at least 10 probes in a CNV call. A preliminary set of 10,865 CNV Events (CNVE) for design of the genotyping array were identified by merging CNV calls across samples (requiring 40% reciprocal overlap).

The 10,865 CNVs in this preliminary set breakdown into the following four classes:

- 3,284 called in 2 or more CEU
- 897 called in 1 CEU and 1 or more YRI
- 2,387 called in 1 CEU and 0 YRI
- 4,297 called in 0 CEU and 1 or more YRI (of which 2,782 are singletons)

Within overlapping CNVs, the longest contiguous region most unique to that CNV was identified as target for probe design (Supplementary Figure 7).

As not all GSV loci could be accommodated on the array, the loci were sorted hierarchically according to a number of criteria so as to prioritise the most common CNVs in individuals with European ancestry, and CNVs most likely to have a functional impact, as assessed by overlap with functional genomic annotation (genes and ultraconserved elements). The only CNVs that could not be submitted for inclusion on the CNV genotyping array are a subset of the singleton CNVs called only in one YRI individual which overlap neither genes, nor ultraconserved elements. Ten probes were designed into each most unique interval of each CNV.

CNV loci: Affymetrix 6.0 chip + Illumina 1M chip Common CNVs mined from two SNP chip datasets were compared against those described above from the GSV CNV discovery project to identify common CNVs that may not be in the GSV map either due to population sampling or incomplete power to detect CNV.

These two CNV datasets comprise:

1. 1,303 clusterable CNVs observed among the 270 HapMap samples in Broad analyses of Affymetrix v6 chip data[104].

2. 2,574 CNVs called in 2 or more individuals from a preliminary analysis of the Illumina 1M data on the ∼1,200 individuals in the HapMap3 sample set (to be described in a forthcoming manuscript)

In both datasets, 92-93% of the common CNVs (minor allele frequency (MAF)>5%) were also seen in the GSV CNV map, suggesting that the GSV CNV map is relatively complete for the larger (>5kb) CNVs that can be captured on these SNP chips.

All 85 CNVs seen in 2 or more of the 60 unrelated CEU individuals, but not in the GSV CNV map, were selected from the Affymetrix 6.0 dataset for inclusion in the array design, with 10 probes per CNV.

All 85 CNVs >500bp in size and seen in 10 or more of the ∼1200 HapMap3 individuals, but not in the GSV CNV map, were selected from the Illumina 1M dataset, with 10 probes per CNV.

CNV loci: Common WTCCC1 CNVs CNVs were called from a preliminary analysis of normalised intensity data from the Affymetrix 500k data collected in the WTCCC1 project (to be described in a forthcoming manuscript). 15,174 samples remained after QC, with a median of 16 CNVs called per sample. In these samples 210,453 individual CNV calls were made and subsequently merged into 23,453 CNVs (requiring 40% reciprocal overlap for merging).

97% (28/29) of CNVs called in 750 or more (∼5%) of the WTCCC1 samples passing QC were observed in the GSV CNV map.

228 (39%) of the 579 CNVs called in 30 or more of the WTCCC1 samples were not observed in the GSV CNV map and were selected for inclusion in the array design. For each of these CNVs, a core region of common overlap between CNV calls was used to define a region of highest confidence for the presence of the CNV, for targeting with 10 probes.

CNV loci: Novel sequence insertions We sought to identify copy number variable sequences not present in the reference sequence, and thus not detectable in array-based experiments predicated on the reference genome assembly. We considered two sources for these novel sequences:

1. Identification of novel sequence insertions from published assemblies of unmapped fosmid end-pair sequences[105]

2. Sequences present in the genome sequence of Craig Venter, but not in the reference sequence[106]

Kidd *et al.*[105] used analyses of fosmid end-pair sequences from 9 individuals (1 Polymorphism Discovery Resource (NA15510), 2 CEU, 1 CHB, 1 JPT, 5 YRI) to identify 525 sites of novel sequence insertions relative to the reference sequence, and subsequently determined by array-CGH that 186 vary in copy number among a small set of HapMap individuals.

We identified sequence contigs corresponding to these 186 copy number variable novel sequence insertions for inclusion in the array design.

Through comparison of the Venter genome assembly with the reference genome assembly (NCBI36), we identified 4,392 sequences longer than 1kb in length that were present in the Venter genome sequence, but not in the reference genome asssembly. We repeat-masked these sequences, and selected 106 insertion sequences with 3-8kb of non-repeat-masked sequence. We reasoned that common copy number variable novel insertion sequences >8kb in length are likely to have been identified by Kidd *et al.*[105], given the resolution of the Kidd *et al.*[105] approach. We also reasoned that given the lack of concrete evidence that these sequences are polymorphic CNVs, it was not worth including more, smaller, novel sequences

at the expense of CNVs from other sources, which are already known to be copy number variable.

Thus, in total, we identified 292 novel insertion sequences for inclusion in the array design, with ten probes to be designed against each sequence.

Candidate loci: genes selected by disease Disease investigators each selected 5-10 genes for targeting on the array design to identify novel, potentially rare highly penetrant CNVs of potential functional relevance to their disease (Supplementary Table 4)

In total, 994 exons in 66 genes were targeted, with 3 probes per exon. On the Agilent platform it was considered that 3 probes should be sufficient for confident CNV discovery (as opposed to the 10 required for robust CNV genotyping).

Validation loci: from WTCCC1 The WTCCC1 CNV analysis group have selected, from the CNV analyses of the Affymetrix 500k data, 26 loci with some statistical support and/or biological plausibility for association with disease susceptibility in the seven WTCCC1 diseases (to be described fully in a forthcoming manuscript). All seven diseases are represented among these 26 loci. These loci represent 3 common CNVs, 17 genes harbouring rare CNVs, and 6 genomic windows harbouring rare CNVs. We reasoned that including these regions on the WTCCC array would allow fine-mapping of the common CNVs, validation of the association signals from common and rare CNVs, and, potentially, the identification of smaller rare CNVs of the same sequences that could strengthen the signal of association based on only the larger variants detectable from the Affymetrix 500k data.

We designed 3 probes per exon for the gene-based analysis, 10 probes per common CNV, and 20 probes per genomic window.

Analysis of the candidate loci and rare WTCCC1 CNVs is beyond the scope of this paper and will be described elsewhere.

3.2 Sample handling

Each participating sample collection was issued unique WTCCC barcode labels and a manifest spreadsheet with unique sample identifiers for logging data on case/ control status, DNA concentration (requested at 100ngμl^{-1}), DNA extraction method, gender and broad geographical region. Each collection supplied 5-10µg aliquots of anonymized samples in bar-coded deep 96-well plates. The majority of the samples used were the original aliquot supplied for the WTCCC1 GWA study[46], the remaining 5462 samples were either the additional Breast Cancer disease cohort (N=2046 (N=1999 blood and N=47 cell-line replicates)), fresh aliquots of the same samples used in WTCCC1 (N=1876) or replacements for samples excluded by WTCCC1 analysis, that could not be resupplied or failed QC after resupply (N=1540).

On receipt, samples had their DNA concentration measured by Picogreen (triplicate measurements), were checked for DNA degradation on a 0.75% agarose gel and were genotyped in multiplex reactions using the MassExtend (hME) and/ or iPLEX assay[107]; these assays were used to obtain a molecular fingerprint and to confirm the gender of each sample. Note that over the course of the project the iPLEX reaction plexes had changed to increase marker density yielding up to 32 SNPs, a cut-off of over 70% of typable SNPs (dependent on number of SNPs per plex) was used as a QC pass criterion. Samples with concentrations $\geq$50ngμl^{-1} (493 samples were used at 25ngμl^{-1}), showing limited or no degradation, that passed the SNP typing QC threshold and had gender markers in agreement or not violating the supplied information were deemed fit for CNV genotyping.

2000 samples were selected from each disease collection and 1500 from each control collection. Selected samples were normalized to 50ngμl^{-1} and re-arrayed robotically into 96-well plates so that each plate was composed of 47 samples in rows E to H with well H12 left blank for the addition of a common female reference sample (NA10847 for the initial 26 plates of the project and then NA12878; this sample was used as a quality control measure for the entire plate). The plates were then assigned a running order that randomized the collections and plates within each collection but provided a phased bias towards obtaining full datasets for some collections before others to allow analysis pipeline development on full-cohort data before the entire screening process was finished. The phasing method is outlined in Supplementary Figure 6. This running order also ensured that no plates from the same collection or any plates from either of the control collections plates were adjacent on the list. Then new plates were robotically arrayed such that samples in rows E and F from adjacent plates in the running order were switched to generate final screening plates that contained a 24:23 sample ratio of 2 different collections on each plate.

3.3 CNV laboratory pipeline

CNV genotyping was performed using the CGH CNV genotyping array (described in section 3.1 above) at Oxford Gene Technology labs (OGT, Oxford, UK) (the same laboratory that had undertaken the Agilent platform analyses in the pilot study described in section 1 above). Sample tracking and QA/QC were coordinated using OGT's in-house Laboratory Information Management System (LIMS). Sample labelling, hybridization and scanning, was carried out following Agilent's Oligonucleotide Array-Based CGH for Genomic DNA Analysis protocol version 5.0 (P/N G4410-90010). The method utilizes an enzymatic methodology to differentially fluorescently label genomic DNA samples. In brief, in parallel aliquots of 500µg of the sample and the reference DNA (a pool of 10 UK Caucasian genomic DNAs (9 males and 1 female) derived from cell-lines in the European Collection of Cell Cultures (ECACC) Human random control collection (see Supplementary Table 6) were digested with *AluI* and *RsaI*, then hybridized with random primers and labeled in a Klenow extension reaction with Cyanine 3-dUTP (reference DNA) or Cyanine 5-dUTP (case/control sample), purified by size filtration and subsequently measured for specific activity and yield before combining the sample and reference DNA. The samples were then blocked with Cot-1 DNA, hybridized to the 2x 105k microarray slides for 40 hours then washed and scanned.

3.3.1 Use of pooled reference DNA

For CGH-based genotyping, a key issue is that the reference DNA is as similar as possible across CGH experiments, which motivates minimizing any batch-to-batch variation in reference DNA. The $\sim$10 milligrams of reference DNA required for our CGH experiments was prepared in a single batch, as a pool of ten UK genomic DNAs. Our motivation for using a pooled reference sample in our CNV genotyping study was to minimise the difficulties we identified in the pilot study associated with robust clustering of CNVs at loci where the reference sample is homozygously deleted. The absolute amount of DNA in the reference pool at any one locus is not important (as long as it is non-zero) as this absolute amount only influences the relative position of the copy number clusters, not their inherent clusterability.

Supplementary Figure 8 shows that moving from 2 chromosomes (a single reference individual) to 20 chromosomes (a pool of 10 reference individuals) results in a significant reduction in the proportion of loci at which the reference pool is homozygously deleted, but that further increases in the pool size result in minimal reductions in this proportion of loci.

3.4 Selection of samples for repeat assays for reasons of data quality

The service contract between WTCCC and Agilent/ OGT for this project made provision for samples to be repeated if they failed agreed pre-defined technical QC metrics. In addition the contract allowed 6.5% of samples to be selected by WTCCC for repeat (i.e. for any reason decided by WTCCC during the project). Using these two routes, a total 1709 samples underwent a repeat assay.

Agreed pre-defined technical QC metrics - All the samples were repeated from plates where at least one of the following occurred: (a) the female control had a derivative log ratio spread (DLRS; a measure of probe-to-probe variation calculated by taking the interquartile range of the vector of differences between the log-ratio measurement at sequential probes along the genome, and scaling this to account for the effects of averaging the noise) greater than 0.2, (b) either the Red or Green channel showed a median background subtracted signal intensity < 50, Background Noise > 10 or Signal to Noise $<$ 30. These criteria resulted in 273 repeats being undertaken. A further 127 samples were repeated because they had failed to meet OGT's strict QC requirements prior to scanning.

Samples selected by WTCCC - Another 1309 samples were selected for repeat based on specifically developed calling-based metrics that used the post-calling posterior probabilities of copy-number class assignment and the dispersion of samples from the median of the assigned copy number class, and Agilent QC metrics, particularly the DLRS which we found to correlate highly with CNV genotype quality and post-calling metrics (data not shown).

Overall, 1287 samples were repeated using 400µg of sample from the original aliquot supplied and then randomly re-arrayed for the repeat phase, 422 samples, selected due concerns over sample mixing or due to outlying red intensity labeling values were resupplied after concentration rechecking and if necessary renormalisation of the samples to 50ngµl^{-1}.

3.5 Duplicate samples

Duplicate samples from within each cohort were added to the study to enable estimation of background noise for each probe and call concordance (Supplementary Table

2). 47 samples from each cohort were randomly selected and arrayed randomly onto new screening plates which were inserted randomly throughout the screening plate running order with the exception of the UKBS cohort were 2 single cohort repeat plates were generated and BC where 47 matched blood and cell-line derived samples were selected and arrayed in pairs across 2 screening plates. An additional 46 duplicates were accidentally added during the plate generation phase, either due to multiple submissions of the samples within the cohort, manifest errors or accidental re-selection. During analysis further duplicates and close relatives samples were identified (section 5.1). There was no information from the manifests that could be used to identify these duplicates before screening.

4 Data and pre-processing

4.1 Introduction: Making sense of array CGH data for CNV calling.

Here we aim to provide some insights into the challenges of genome-wide CNV typing in large samples. The statistical problems here are analogous to those of calling genotypes from SNP-array data. In our experience they are much more challenging, for several reasons.

Each CNV on the array is targeted by a set of probes (usually, but not always, 10 probes per CNV see Supplementary Figure 9 for a histogram of probe numbers per CNV). The experiment measures the intensity of each probe after binding to the test sample, and after binding to a paired reference sample. The ratio of these two intensities then provides a surrogate for the relative amount of DNA in the test compared to the reference sample. In practice, functions of this ratio, such as its logarithm, can have better statistical properties. Our standard pipeline added a constant to the ratio before taking logarithms to ameliorate some numerical instabilities.

Some form of normalisation is usually applied to the intensity data. The aim of normalisation here is to minimize differences between samples. As a simple example, if the amount of DNA applied to the array were greater for sample A than for sample B, then intensity measurements for sample A would tend to be bigger than those for B even when both samples had the same number of copies of the CNV in question. So-called quantile normalisation is a procedure which forces the distribution of all the intensities measured for sample A to match those for all the intensities for sample B (typically by matching both to a third, common, distribution). Normalisation could be applied separately to each of the test and reference samples in a pair, or to their ratio, or to the logarithm of their ratio. We explored many such choices, with no single choice working best for all CNVs. Indeed, as we note below, there are some CNVs where the signal is only clear without any normalisation.

Supplementary Figure 10(a) shows a histogram of normalised log relative intensity across all samples in our experiment for each of the probes in a particular CNV (CNVR3337.4). Note that some probes clearly exhibit variation consistent (here) with three copy-number classes, while others do not, appearing instead to display noise rather than useful signal. This is not uncommon in our data. While visual inspection of each probe-histogram for a particular CNV could be used to choose a subset of probes to take forward for analysis, such manual curation is not practicable genome-wide.

Two natural, automated, approaches to combining the information across probes in a single CNV are (i) to take the mean across all probes, and (ii) to take the first principal component (PC). The mean is a natural summary. The first PC provides the linear combination of the individual probe measurements with the greatest variation across samples, which would hopefully coincide with real variation in copy-number at that CNV. We found the first PC tended to perform better in our data than the mean. Supplementary Figure 10(b) shows the histograms, for three of our 10 collections, of the intensity for each sample as summarised by the first PC across probes for the same CNV (CNVR3337.4). Note the continuing lack of a clear signal.

We developed an automated approach to picking the best probes per CNV by taking advantage of the duplicate samples in our experiment. Where a probe is measuring a real biological signal, it should give similar values across replicates of the same sample. On the other hand, where a probe is largely measuring noise, it should show greater variation across replicates. For each probe on the array, we estimated its variance across replicates of the same sample from the set of duplicates in our experiment. Probe variance scaling (PVS) then combines information across probes within a CNV (for either their mean or first PC) by weighting probes inversely with their estimated variance across replicates. Supplementary Figure 10(c) shows the application of PVS to CNVR3337.4, revealing a clear signal for two copy-number classes. If, in principle, one knew the correct assignment of samples to copy-number classes, then it would be possible to weight probes in such a way as to maximise discrimination between these classes (this is a standard statistical problem solved by a technique called linear discriminant analysis). Although we do not know, a priori, the correct assignment, this principle motivates an iterative procedure: use some sensible approach to assign individuals to copy-number classes, and then chose probe weightings to maximise discrimination amongst these classes, and then re-cluster into classes 25. We adopted a single-step version of this procedure, denoted by LDF (linear discriminant function).

The most successful one-size-fits-all approach to analysis of our data involved a particular choice of normalisation, combination of probes via PVS and the first PC, followed by LDF. Under this approach (called our standard pipeline) we obtained good data for 2,716 CNVs. (Without PVS and LDF that number drops to 1,925 CNVs.) But a key finding is that no single approach works well for all CNVs. As a consequence we carried forward 16

different analysis pipelines, and then chose the pipeline which yielded the best calls, as measured by QC criteria. Collectively, these 16 pipelines yielded good data at 3,432 CNVs. Of these, the standard pipeline provided the best data for 524 CNVs (15

To illustrate some of the points above, Supplementary Figure 11(a) gives an example of a CNV with much clearer signal using a normalisation different from our standard one (in this case no normalisation). Supplementary Figure 11(b) illustrates a situation where using the mean as a probe summary performs much better than does using the first PC.

Even prior to any of the analyses described above, there is a challenge in deciding which probes to use for a particular CNV. To illustrate this, imagine a situation where a small deletion occurs inside a larger deletion. To assay the internal deletion it would be natural to use the probes inside that deletion. On the other hand, when assaying the larger deletion it will typically be more efficient to use only the probes not in the inner deletion, to avoid confounding of the two signals. These situations do occur, though fortunately not frequently. We investigated various sophisticated statistical procedures for automating the solution to what we call this probe-to-CNV mapping problem, but settled on a relatively simple procedure in our final analysis (see next section for details). As with all of the challenges described in this section, careful manual examination and curation of the probe-level data can often improve on automated procedures for a particular CNV of interest, but this is impracticable on genomic scales.

4.2 Probe-to-CNV mapping

The WTCCC CNV genotyping array contains a total of 105,072 features, consisting of 102,602 unique probe sequences including those designed to target CNV loci (96,959), novel insert regions (2,907) and exonic regions for genes of interest (2,633), together with 103 Agilent control probes distributed throughout the array. Many of the 103 control probes are replicated multiple times on the array and so there are 1060 oligos on the array that are Agilent controls. Each of the unique probes was assigned a unique index value for cross referencing of the feature-level data. The sequence of each of the unique probes on the CGH CNV genotyping array was aligned to the reference human genome (build 36) using the BLAST-like Alignment Tool (BLAT)[108] to identify all genomic loci having 100 % sequence similarity, producing a list of multiple genomic coordinates for each probe sequence present on the array.

In the intervening time between the design of the CNV array and the analysis of the resulting CNV genotyping data, the Genome Structural Variation Consortium underwent further analyses to improve the definitions of CNV breakpoints from the 42M discovery data. It was decided to use the latest version of the GSV CNV definitions to ensure concordance between the two studies. In particular, this was done to ensure that CNV identifiers match between the present study and the GSV publication of Conrad *et al.*[103]. The CNV genotyping array was originally designed to target a total of 9,722 CNVs from the version 1 GSV 42M CNV set of 10,865 CNVEs. Using the version 2 GSV 42M CNV set of 11,700 CNVEs, we identified 10,835 CNVs showing overlap of at least 1 bp with probes on the array.

A map between the probes on the array and the updated set of 12,740 targeted loci (11,246 CNVs, 918 PI selected exonic regions of interest, 274 exonic validation regions from the original WTCCC1 association study[46], 292 novel insert sequences and 10 X-chromosome non-polymorphic control regions) was generated by finding all probes with at least one set of genomic coordinates showing an overlap of 1 base pair or greater with the targeted locus. However, regions of polymorphism containing multiple overlapping CNV events, together with the effect of probes that may detect signal from multiple distinct genomic regions due to multiple alignments to the reference sequence, can lead to difficulties in ensuring that the reported signal for each locus truly represents the copy number state at this site and is not polluted by the signal from overlapping events. To correct for this, we took only the subset of the probes intersecting the targeted locus which map to the least number of additional loci, ensuring that probes used to report signal represent the most unique region of the locus. This algorithm acted as a filter, removing 10,819 probes which map to an unusually high number of potentially polymorhic loci.

Supplementary Figure 9 shows the distribution of the number of probes mapped to each locus, and Supplementary Table 5 shows the number of loci (for each distinct class) for which different numbers of unique probes were found to map. As described in Section 3.1, 3 probes were chosen to target exons and 10 probes were chosen to target other loci, and prominent peaks for these two values can be seen in Supplementary Figure 9. For the vast majority of exons and novel inserts (90.4 % and 99.3 % respectively), the probes designated as targeting the loci following mapping were specifically those designed in the first instance. However, only 55.2 % of the CNV

loci targeted on the array were found to overlap specifically with the 10 probes originally designed to the region. This is due to a combination of factors, including removal of probes that map to multiple loci by the algorithm as described above, the effect of probes mapping to multiple genomic coordinates, and overlap between loci in complex regions of the genome. There were also some loci where the expected number of probes could not be successfully designed, in which case fewer than 3 or 10 probes were used and replicate probes were included to make up the number. Also, moving to the latest set of GSV CNV definitions as described above meant that, for many CNVs, the reported breakpoints were altered from those used for the array design, resulting in a different subset of probes being identified in the mapping stage. For the majority of the 5,308 CNVs not targeted by exactly 10 probes (81.3 %), this was due to a smaller number of probes than expected being mapped to the CNV locus, suggesting that the filtering of probes implicit in the mapping algorithm was the primary source of discrepencies in probe numbers. However, there were a small number of loci for which a larger number of probes than expected were detected in the mapping step, which typically relate to loci on the larger side of the size spectrum.

The resulting map was used to generate feature-level data for each locus, containing signal data for all features whose probe sequence was found to map to a location with at least one base pair in common with the locus. Signal values for probes repeated multiple times on the array were combined for each sample by taking the mean. These per-locus probe-level data were summarized to give a univariate measurement for each locus as described below (Section 4.4).

4.3 Data Normalization

Red channel (test DNA) and green channel (reference DNA) data were further processed by OGT using Agilent Feature Extraction software version 9.5.3.1 [109]. First, local background signal was removed for each feature. Features registering a non-zero signal, or a signal that was not significant versus the background (based on an additive error model) were adjusted to avoid biases in the log-ratio calculations. Next, a multiplicative detrend algorithm [109] was applied across all features on the array to correct for linear variations in the intensity values due to non-homogeneity of hybridisation. Finally, a per-channel dye-bias linear normalization method was applied, which set the geometric mean of the signal intensity to a constant value of 1,000 for all non-control probes with significant signal versus the background.

Data were further normalized in-house, and a number of normalization schemes were considered to identify the optimum method to ensure maximum genotyping potential. However, we found that the optimum normalization was highly dependent on the properties of the individual loci. Data were normalized in each of the following ways, and the optimum normalized data set was selected on a per-locus basis:

- Red channel data only – R
- Quantile-normalized red channel data only – $\mathrm{QNorm}(R)$
- Green channel data only – G
- Quantile-normalized green channel data only – $\mathrm{QNorm}(G)$
- Ratio of the red and green channel data – R/G
- Quantile-normalized ratio of the red and green channel data – $\mathrm{QNorm}(R/G)$
- Log_2 ratio of the red and green channel data – $log_2(R/G)$
- Quantile-normalized log_2 ratio of the red and green channel data – $\mathrm{QNorm}(log_2(R/G))$
- Log_2 ratio of the quantile-normalized red and green channel data - $log_2(\mathrm{QNorm}(R)/\mathrm{QNorm}(G))$
- Quantile-normalized log_2 ratio of the quantile-normalized red and green channel data - $\mathrm{QNorm}(log_2(\mathrm{QNorm}(R)/\mathrm{QNorm}(G)))$
- Transformed log_2 ratio of the red and green channel data - $log_2(R/G + 0.5)$
- Transformed log_2 ratio of the quantile-normalized red and green channel data - $log_2(\mathrm{QNorm}(R)/\mathrm{QNorm}(G) + 0.5)$

Univariate quantile normalization [110] was performed for each sample using the NormTools package of C++ normalization and preprocessing tools, which are freely available from `http://cnv-tools.sourceforge.net/Normtools.html`. Quantile normalization was performed seperately for males and females for all probes identified as aligning to the X-chromosome (see below). The reported gender of each sample was checked by performing

genotyping using CNVtools[5] on 10 non-polymorphic X-chromosome control loci. A mixture model was fitted with two components, and the resulting copy-number assignment was used to assign gender calls for each sample (1 males, 2 females). If the copy-number calls disagreed with the reported gender for greater than 4 of the 10 control loci, the reported gender was taken and the sample was flagged for future QC. The feature-level data for each sample were split into two distinct subsets; those with at least one alignment to the X-chromosome, and those with no alignment to the X-chromosome (including control probes and novel insert probes). The X-chromosome probe-level data were further split, with male and female individuals quantile normalized in distinct groups. Univariate target distributions were generated for each group by taking the mean of the ordered univariate signals over a random subset of 1,000 good quality samples (distributed evenly between all cohorts). Univariate quantile normalization was performed on a per-sample basis for each group by adjusting the univariate signal distibution such that the ordered signal values match those of the target distribution.

4.4 Probe summaries

In order to obtain reliable CNV genotype calls, multiple probes were designed within each targeted CNV region (see Section 3.1). However, both calling algorithms used in this study require a one-dimensional data summary for each sample. Therefore, for each CNV, the signal at multiple probes was summarised to obtain a single value for each sample.

After rescaling the probes (see Section 4.5), we used the probe summary procedure proposed in ref. 5. This approach first uses a principal component analysis (PCA), taking the first principal component to summarise the data across probes. These were used to generate an initial set of calls using the calling algorithms. The calls are then used in a canonical correlation analysis (ref. 111, chapter 4) to find the linear combination of probes maximally correlated to the posterior probability of genotype calls. This procedure can be interpreted as an approximate maximum likelihood estimation procedure, via the EM algorithm, for an extended mixture model in which the signal is multidimensional (Plagnol, Clayton, Complex Disease Association studies, chapter 14, forthcoming).

4.5 Probe variance scaling (PVS)

The relative scales of the individual probes will affect the PCA step in the summarisation procedure (see Section 4.4), with probes that have higher variance having a greater influence on the resulting probe summary. If probes are well-calibrated, higher variance is expected for probes that measure variable copy number, which is desirable. However, higher variance can also result from elevated background noise, in which case such probes should be downweighted. To determine an appropriate scale for each probe, we used the duplicate samples in a procedure we refer to as probe variance scaling (PVS).

Firstly, from all of the duplicate samples we selected a subset that were high-quality, by taking only samples initially designed as duplicates (i.e. not including samples repeated for other reasons) and having a correlation of at least 0.8 between the corresponding duplicate pairs on an initial set of CNV calls. This gave 427 pairs of duplicates.

Secondly, we estimated the within-sample standard deviation for each probe using measurements from these duplicates, as follows. For a given probe, let x_{i1} and x_{i2} be the two duplicate measurements for the ith individual, and let there be n pairs of duplicates in total ($n = 427$ for our data). The standard deviation is estimated by,

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(x_{i1}-x_{i2})^2}{2}}\,.$$

This is equivalent to a within-groups standard deviation in an ANOVA using n groups of 2 observations each.

Finally, we rescaled each probe by this value (after centering), to obtain probes with identical within-sample probe variances. In particular, letting x_j be the probe intensity for the jth sample and m be the mean across all samples, the rescaled intensity is,

$$x'_j = \frac{x_j - m}{s}\,.$$

These rescaled values were then used in the PCA step of the probe summarisation procedure.

We found that PVS gave a substantial improvement in cluster separation for a large number of CNVs. Removing the PVS step from the probe summarisation and calling procedure resulted in 460 fewer well-separated, polymorphic CNVs (see section 5.2 for definition of well-separated, polymorphic CNVs).

Figure 1 from box 1 of the main paper shows an example of a CNV where PVS is crucial in extracting the signal of polymorphism.

5 Quality control procedures

5.1 Sample quality control filters

Two sample exclusion lists were constructed and used in the analysis of the data. The first list (pre-calling exclusion list) was used to exclude samples from the final calling of the CNVs using the processed intensity data. The second list (pre-testing exclusion list) was used to exclude samples from the testing for CNV association based on the final set of CNV calls. A full break down of excluded samples is given in Supplementary Table 8.

Pre-calling exclusions

1963 samples were excluded from the final CNV calling based on several different criteria described below. Some of the filters were applied to the raw intensity data while others were based on CNV calls obtained from an initial calling run on the data.

Supplier error 149 samples were excluded due to evidence that the samples were not the same as those indicated by the supplier manifest. Sequenom QC and calling gender on the CNV array were used to confirm these discrepancies.

Sample handling error 15 samples were excluded due to evidence of an error during arraying the samples for CNV screening.

Multi-cohort duplicates 18 samples (9 pairs) were detected that showed high correlation with another sample from a different cohort, indicating a sample that has genuinely been collected twice as the patient has at least two of diseases. No sample handling issue could be detected, and the data matched for both samples with the Sequenom and WTCCC1 SNP data. Both samples in the pair were excluded. The samples were identified by taking the summarised probe-level signal (first principal component) over 1,500 good quality polymorphic CNVs and running an all-vs-all correlation analysis (Pearson) to identify highly correlated samples.

Non-European samples 26 samples were excluded due to evidence of non-European ancestry. A PCA analysis was carried out on CNV calls from an initial calling run, that included HapMap individuals from the CEU, YRI and JPT+CHB panels. Examination of the loadings and scores of this analysis indicated that only the first principal component was discriminating European samples from the YRI and JPT+CHB samples. Supplementary Figure 12 shows the scores for each sample from the first principal component and highlights 14 outlying BC samples that were excluded. A further 11 CD samples and 1 RA samples were also excluded based on self-reported ancestry information.

Mixed sample 189 samples were excluded due to the samples having a high correlation with another sample on the same well of the screening plate pair or an adjacent well in the same plate suggesting that these samples consist of a mixture of DNA from two or more non-identical individuals.

Low signal 72 samples were excluded due to having a low signal intensity for either the green or the red channel (< 100). The precise quantities used are the metrics named "SignalIntensityRed" and "SignalIntensityGreen" from the Agilent Feature Extraction software [109]. These give a measure of the median background-subtracted red and green channel signals respectively (not logged) across all non-control probes on the array.

High derivative log ratio spread Samples were excluded based on a measure of the variability in log-ratio ($\log_2(R/G)$) across all probes for each sample. The Agilent DLRS metric was used which is measures the spread of the differences between the log ratio values of consecutive probes [109]. High values of this metric indicate a poor sample. We excluded samples if DLRS was either > 0.35, or > 0.3 if it is a repeat and the original sample had a DLRS > 0.35.

Outlying CAD samples 405 CAD samples were identified that noticeably reduced the ability to distinguish different CNV classes when the samples were included. Removing these samples lead to a clear improvement in the ability to cluster some CNVs in the CAD cohort. This problem was observed for multiple probes in this study and is illustrated in Supplementary Figure 13 (see first and second panels) where we extracted from CNV ILMN_1M_4 a subset of probes (A_16_P30155705, chr1_047654910_047654955, A_16_P30155706, chr1_047654921_047654966, chr1_047654923_047654968, A_16_P30155708) that showed no sign of CNV polymorphism in the non CAD cohorts. However, a set of CAD samples was clearly separated from the main distribution at these probes.

To identify the subset of problematic CAD samples we used two probe sets (average signal for ILMN_1M_4 probes described above and probes A_18_P20232231, A_16_P40333900, A_16_P02994736 in CNV CNVR6314.1) outside of CNV regions for which the separation of outlying CAD samples was particularly obvious. For both probe sets, we manually set cutoffs for the mean normalized signal value and we excluded samples that exceeded both cutoffs (see the third panel of Supplementary Figure 13 with excluded samples marked in red).

Further analysis of the processing pipeline indicated that the likely source of the problem was mis-calibrated DNA concentration. Variable DNA concentrations differentially affected each probe, thus altering the within sample probe intensity rankings. In quantile normalisation, probe intensities were first ranked within the sample, and each intensity data point was then replaced by the appropriate quantile of the marginal distribution of probe intensities over all samples. Therefore, altered probe rankings eventually affected the normalized signal distribution.

Initial-calling quality metric 409 samples were identified based on 3 metrics designed to measure the quality of samples from an initial set of calls. The three metrics were (a) average CNV call rate measured as the proportion of CNV calls made on each sample using a calling threshold of 0.95, (b) average posterior probability of the most likely CNV class across all CNVs for a sample, and (c) average log-density (from the final model fit after merging) across all CNVs for a sample. Samples were ranked according to the minimum of the ranks on these three metrics and sample excluded so that the total number of exclusions was 2% of the total sample size.

Pre-testing exclusions

A further 1832 samples were excluded before testing for association of CNVs with the disease phenotypes. This resulted in a total of 17304 samples used in testing.

Post-calling quality metric 1099 samples were excluded based on thresholding three metrics applied to a final set of calls from the CNVCALL and CNVtools standard calling pipelines.

Dispersion metric A set of hard calls were made using CNVtools. A hard call is the genotype with the maximum likelihood given the estimates of the model parameters. For each CNV these hard calls were used to generate empirical means and standard deviations of the components that individuals were assigned to (the sample means conditional on the calls). Then for each individual at each CNV the absolute distance from the mean of the distribution that individual was assigned to was calculated. These were then averaged across CNVs to get the dispersion statistic for each individual. A threshold of 1.3 was chosen after visual inspection, all individuals that exceeded this threshold were excluded from testing (see Supplementary Figure 14).

Posterior Probablistic calls were made at each CNV using CNVCALL. For each individual the probability of assignment to the most-likely (non-null) class was averaged across all the CNVs polymorphic after merging. A threshold of 0.967 was chosen after visual inspection, all individuals that failed to exceed this threshold were excluded from testing (see Supplementary Figure 15).

Heterozygosity Using hard-calls from the CNVCALL (thresholded at a value of 0.95) the proportion of heterozygote calls in each individual was calculated on the CNVs polymorphic after merging. As this is a sum of independent binomials the Central Limit Theorem Applies. Modelling this as a normal distribution using the median as a robust estimator of the mean of the distribution, individuals were excluded if they lay in either tail with the probability of exclusion set at 1/2000 under the null (see Supplementary Figure 16).

Duplicates and close relatives 734 samples were excluded because they were identified to be duplicates or closely related samples. Samples from the same individual (duplicated samples) were identified as those having a calls correlation (using hard calls at a 0.95 threshold) of > 0.9. Closely related samples were identified as those having a calls correlation of between 0.6 and 0.9. Supplementary Figure 17 shows a plot of maximum calls correlation for each sample with any other sample. For each set of samples from the same individual, only the sample with the highest average posterior was retained. Likewise, for closely related samples from the same collection, only the sample with the highest average posterior was retained.

5.2 CNV quality control filters

We used 16 different analysis pipelines where different aspects of the data pre-processing were varied. Sup-

plementary Table 9 gives the details of these pipelines. To choose which pipeline to use for a given CNV we used the pipeline which gave the highest number of classes, and the highest average posterior in cases where more than 1 pipeline gave the same maximum number of classes. At CNVs on the X chromosome only CNV calls in females where used when apply CNV QC filters. 6,568 CNVs (59%) were called with just one class and these were removed from further an analyses. Of the remaining 4,539 CNVs, a further 894 (20%) were removed as they had poor post-calling quality metrics (average posterior < 0.98). In addition, where discovered CNVs overlapped, we were able to use our experimental results to determine whether they were genuinely different. Where overlapping CNVs had highly correlated calls ($r^2 > 0.995$), we assume they correspond to the same CNV. In such cases we chose to analyse the CNV with the highest average posterior across samples. Of the 3,645 CNVs that had good separation 213 were removed as they were highly correlated with an overlapping CNV. Testing analyses thus focussed on 3,432 CNVs. A table of basic data for each CNV is available from `http://www.wtccc.org.uk/wtcccplus_cnv/supplemental.shtml`.

6 Calling and testing

Two CNV calling algorithms (CNVtools and CNVCALL, see sections below) were used. The estimated numbers of classes for both calling algorithms for the 3,339 autosomal CNVs passing QC are shown in Supplementary Table 7

6.1 CNVtools

Principle

The association testing approach has been described previously[5] and here we sketch the principle. We used a likelihood ratio approach to test for association between the genotype calls and the case-control status. Genotypes are called using a finite mixture model. Formally, this association test can be as summarized as jointly fitting two linear models:

$$X = \gamma + \theta^t Z + \epsilon \quad (1)$$

$$\text{logit}(Y) = \alpha + \beta X \quad (2)$$

- X is a N-dimensional vector of signal intensities, where N is the number of samples in the study.
- Z is the (N, G) matrix of genotype assignment, where G designates the number of copy number classes: $Z_{i,j} = 1$ if and only if the sample i has genotype j. Each row z_i of Z is sampled from a multinomial distribution with probabilities $(\Phi_i)_{i=1}^G$ representing the genotype frequencies in the sampled population.
- The error term ϵ is normally distributed with mean 0. Our default assumption is that the standard error σ_{X_i} is a function of the genotype X_i. However, we used a T-distribution model when estimating the number of genotype clusters (see details below).
- θ is a G dimensional vector, linking the genotype status with the mean value of the signal intensity. To protect against differential bias we assume that θ is different for each cohort (see below).
- α and β are scalar and $\beta \neq 0$ under the alternative $\mathbb{H}_1$. Our default assumption is that the log-odds ratio is proportional to the genotype X.
- Y is the N dimensional binary vector describing the case-control status.

Equation (1) describes the clustering model for the genotype calls, and (2) is the traditional Cochrane-Armitage test[112]. The log-likelihood is a function of the parameters $(\gamma, \theta, \alpha, \beta, \sigma)$. To test for association the likelihood is maximized under the null $\mathbb{H}_0 : \beta = 0$ and under the alternative $\mathbb{H}_1 : \beta \neq 0$. The resulting loglikelihood ratio $\Delta L = \log(L_0) - \log(L_1)$ is asymptotically distributed as χ^2 with one degree of freedom. This class of association test is implemented in the `R` package CNVtools[5].

Accounting for differential bias

Several artifactual biases, in particular associated with DNA handling, storage, or genotyping can affect the measured signal intensities of the CNV probes. Therefore, assuming that the clustering model parameters are identical across different case or control collection can inflate the false positive rate[89]. To protect the association test against differential bias, we modified equation 1 as follows:

$$X = \gamma + (Y \star \theta)^t Z + \epsilon \quad (3)$$

where $(Y \star \theta)$ indicates that different vectors θ are used for each collection. For case collections that consist of a combination of blood derived and cell derived DNA, different vectors θ were used for each subset. However, we assumed that the variances are only function of the copy number, and not different across collections.

For CNVs that could not be fitted otherwise we modified the clustering model and added the additional assumption that the variances are equal across copy number to improve the robustness of the clustering. Note that for the simplified model we let the variances differ across collections.

Estimation of the number of components

To estimate the number of components we fitted the model described above six times assuming between one and six genotype groups. To select the most appropriate number of genotype groups we compared the estimated likelihoods for each of the six models using a Bayesian information criteria (BIC, see O'Hagan and Forster,[113] chapter 7).

Visual inspection of the estimated number of components showed that we obtained a more reliable estimate of the number of components by replacing the Gaussian error model with a T-distribution model. However, this estimation step was still relatively inaccurate and for about

4,000 CNVs identified by the automated algorithms as being non monomorphic we manually curated the data and set the number of components after visual inspection of the histogram of data intensity. A R graphical user interface was used to facilitate this manual curation.

Measuring clustering quality

We defined a clustering quality score (Q) that compares the distance between clusters with variation within clusters[5], calculated using the most likely call for each sample. Empirical evidence suggests that when this score is less than 4 the statistical power to detect association drops sharply and we used this threshold to exclude poorly clustered CNVs[5].

Association testing for X linked CNVs

X linked CNVs need to be treated differently when testing for disease association because the distribution of the number of copies necessarily differ between men and women. When analyzing di-allelic CNVs or SNPs, assumptions can be made to obtain a test statistic distributed as 1 degree of freedom χ^2 under the null of no association (see Ref 114). However, in the context of potentially multi-allelic CNVs it is not clear what assumptions should be made. Therefore, we opted for the slightly less powerful, but also more general approach, of separately testing for association in the male and female samples. We obtained two test statistics, each distributed as one degree of freedom χ^2 under the null. We then combined both statistics into a single two degrees of freedom χ^2 statistic under the null hypothesis of no association.

6.2 CNVCALL and CNVTEST

We have developed a Bayesian Hierarchical Model to call CNV genotypes applicable to multi-cohort studies similar to the CHIAMO SNP genotype calling method[46]. The main idea behind this approach is to allow for differences in CNV intensity distributions that may exist between cohorts due to differences in DNA source or DNA storage and handling. The advantage of the hierarchical model is that it allows information to be pooled across cohorts. For example, when there are no differences between cohorts the model effectively calls all the data as one cohort. We provide a brief description of our method here, which will be described in more detail in a subsequent publication[115]. Our strategy involves two steps. In the first step we make probabilistic CNV calls in all of the cohorts together using the hierarchical model. This step is implemented in R and called CNVCALL. We then feed these probabilistic calls into a CNV testing program (CNVTEST) to calculate a Bayes Factor for association of the CNV calls with the disease phenotypes.

Model fitting and making probabilistic CNV calls To fit the model for a given number of classes (K) we use a stochastic search algorithm to maximize the posterior distribution. We then use the *maximum a posteriori* (MAP) estimate of the parameters $(\hat{\theta})$ to calculate the posterior probabiility of CNV class for each individuals, $q_{jik} = P(Z_{ji} = k|\hat{\theta})$, where Z_{ji} is the genotype of the j^{th} individual in the i^{th} collection.

Choosing the number of components It can often be difficult to predict in advance how many classes a CNV locus will exhibit in a given sample of individuals so this must be estimated as part of the CNV calling procedure. We use a procedure based on the Bayesian Information Criteria (BIC)[113] to choose the value of K. Observations from real and simulated data suggest that this performs well. We also have observed that our use of an outlier class noticeably improves the performance of this model selection step.

Merging classes At many CNVs we observed distributions which exhibited skewed distributions. The extent of the skew is heavily influenced by normalisations, transformations and other pre-processing steps applied to the data. We have observed that this can elevate the estimated number of components chosen by BIC. To ameliorate this problem we allow adjacent clusters to *merge* whenever the merged component results in a uni-modal distribution. To identify the merged components that specific individuals lie in we add the posterior probabilities of lying in the pre-merged components that they are comprised of.

Association testing : Calculating Bayes Factors and p-values at CNV loci

To test for association at each CNV locus we use the CNV call probabilities to calculate a Bayes Factor between a null model and a model of association for an additive effect of CNV copy number. We define Φ_{ji} to be a binary case-control phenotype and let M_0 be a models of no association and M_1 be a model of association. Then the Bayes Factor between the two models

is the ratio of the marginal likelihoods of the two models : $BF = P(data|M_1)/P(data|M_0)$. The marginal likelihood for the model M_1 is given by

$$P(data|M_1) = \int P(\Phi|\beta)P(\beta)d\beta, \qquad (4)$$

where

$$\begin{aligned} P(\Phi|\beta) &= \prod_{i=1}^{C}\prod_{j=1}^{N_i}\sum_{k=0}^{K-1} P(\Phi|Z_{ji}=k,\beta)q_{zij}, \\ P(\Phi|Z_{ji}=k,\beta) &= p_{jik}^{\Phi_{ji}}(1-p_{jik})^{1-\Phi_{ji}}, \\ \log\frac{p_{jik}}{1-p_{jik}} &= \beta_1+\beta_2 k, \end{aligned}$$

and C is the number of cohorts, N_i is the number of samples in cohort i and K is the number of copy number classes at the CNV. We use normal distribution priors on the association model parameters $\beta_1 \sim N(0,1)$ and $\beta_2 \sim N(0,0.2^2)$ and use a Laplace approximation to carryout the required integral and estimate the marginal likelihood. The calculations for model M_0 are the same, except that $\beta_2 = 0$ and no prior on this parameter is needed.

To calculate p-values we maximize the likelihood $P(\Phi|\beta)$ under M_1 and M_0 and calculate a maximum likelihood ratio test statistic which can be used to calculate a p-value.

6.3 Plots of CNVs showing evidence of association

A complete set of cluster plots for each of the loci included in Table 3 of the main paper are included in Supplementary Files 2 and 3. Here we provide two example plots for CNVR2523.1 produced by the CNVCALL/CNVTEST approach (Supplementary Figure 18) and the CNVtools approach (Supplementary Figure 19) respectively.

7 Properties of the CNV calls

7.1 SNP tagging

In order to determine how well-tagged the CNVs analysed in our experiment were by SNPs, we carried out correlation analyses using control samples that were common to the current studies and other WTCCC studies. We analysed three different collections of SNPs. We used imputed SNP calls from the WTCCC1 study which used the Affymetrix 500k array, and actual calls from the WTCCC2 study using both the Affymetrix 6.0 array and a custom Illumina 1.2M array. In all cases we used samples from the UKBS collection.

For the WTCCC1 data, we used calls created with CHIAMO[46] and used IMPUTE version 1[116] with release 22 of HapMap CEPH to impute genotypes for all HapMap SNPs. For the WTCCC2 Affymetrix 6.0 data, we used CHIAMO to create SNP genotypes. For the WTCCC2 Illumina data, we used Illuminus[117] on the UKBS samples that had a SNP calls correlation of $>$0.95 with WTCCC1 samples to determine genotype calls.

For all analyses we created "hard calls" using a 0.95 threshold on posterior of both CNV and SNP calls. We removed all SNPs that had only one class in the common UKBS samples, or had missing calls for at least 10% of samples.

For each CNV we calculated the Pearson r^2 value between that CNV genotypes and SNP genotypes. We used only those SNPs within 1MB of the estimated start and end points of that CNV. The use of r^2 directly on genotype data makes possible correlation analysis for CNVs with more than 3 classes. Further, other methods for estimating correlations which directly or indirectly estimate phase information would be inappropriate for any CNV with more than two variants, including some three-class CNVs where both duplications and deletions are present.

We selected the SNP with the highest r^2 values for each CNV. In the case of ties, the first SNP in the list would be chosen, which would be the SNP with the lowest chromosomal location.

We provide SNP tag information separately for different platforms, but for analyses in the main text we combined information and used the best tagging SNP across the three collections of SNPs. Supplementary Figure 20 shows a histogram of maximum correlation across the three collections of SNPs between each CNV and a SNP. A table of SNP tag data for each CNV is available from `http://www.wtccc.org.uk/wtcccplus_cnv/supplemental.shtml`.

7.2 Association analysis of SNPs tagging CNVs

One of the main questions that we are trying to ascertain in this experiment is the extent to which CNVs may account for susceptibility to disease. One question that we may ask of our data is whether or not SNPs that are suspected to have some association with disease may in fact be proxies for some underlying causal CNV. Our analysis of SNPs that are in high LD with CNVs (see Section 7.1) identified high levels of LD with SNPs for many of our CNVs (roughly 68 % CNVs with $MAF > 10\%$ have $R^2 > 0.8$ with at least one SNP). This offers the possibility that perhaps analyses of SNP disease-association may indirectly assess CNV effects.

To analyse the proportion of disease-associated SNPs which may owe their effects to underlying CNV, we looked at the levels of LD between published association SNP loci and CNVs. 103 index SNPs were identified for 98 published loci for the three cohorts with the highest number of validated associations (35 CD, 43 T1D and 20 T2D)[118,119,120,121,46,122,123,124,125,126,127,128,129,130,82 58,131,85,132,133,84,134,83,135,136,137,138,139,140,141,142,143,144 145,146,147,148,149,150,151,152,153]. For each SNP, genotype calls were obtained – either directly genotyped from the Affymetrix v6.0 or Illumina 1.2M CCC2 data if available, or imputed from Affymetrix 500k CCC1 data if not. For cases where 2 or more index SNPs were available for a given locus, the highest R^2 value was taken. Of the 98 loci considered, genotype data were avalable for 95, with no genotypes available across the three platforms for index SNPs for the loci *NOD2*, *CCR5* and *CTLA4*. Direct genotype data were available for all but two of these 95 loci, so for *TNFAIP3* and *CDC123/CAMK1D* index SNPs, imputed genotype calls from CCC1 data were used. For the 3,339 autosomal CNVs that passed our QC filters (section 5.2), R^2 values were calculated between CNV and SNP genotype calls.

Supplementary Table 11 shows the number of loci that show high LD with at least one of the WTCCC CNVs that passed QC, based on index SNP genotypes from three distict data sources. As can be seen, few of the tested loci show high LD between SNPs and CNVs, with only 2 CD-associated SNPs found in LD with at least one CNV ($R^2 > 0.5$).

The first of the two CD-associated SNPs was rs11747270, which is in high LD with two CNVs, CNVR2646.1 ($R^2 = 0.92$ – Affymetrix v6.0) and CNVR2647.1 ($R^2 = 1$ – Affymetrix v6.0), which are the two CD-associated CNVs identified upstream of the *IRGM* locus (described in more detail elsewhere). The second of the CD-associated SNPs was rs2301436 which

shows fairly high LD with CNVR3164.1 ($R^2 = 0.74$ – Illumina 1.2M), a CNV that lies in the intronic portion of the long transcript variant of the chemokine (C-C motif) receptor 6 *CCR6* gene. *CCR6* is the receptor for the β-chemokine *CCL20* which is expressed in epithelia from colon and other intestinal tissue, so presents a possible causal variant for inflammatory bowel disease such as CD. CNVR3164.1 has indeed been found to show some association with CD in our analyses, however the association statistics failed to pass our stringent thresholds for CNV association ($p - value = 2.90 \times 10^{-3}$, $log_{10}(BF) = 1.468$).

These analyses suggest that these three SNP associations may in fact be measuring an underlying disease-causing copy number variant at these loci, although this does not appear to be a widespread occurence. Further analyses are required to make any firm assertions as to the proportion of SNP associations that may relate to CNV, although these data suggest that it may not be a widespread occurence.

7.3 Calculating minor allele frequencies

For CNVs that were called with either 2 or 3 classes, we estimated the underlying minor allele frequency (MAF) assuming it was biallelic. For 2-class CNVs, we assumed the more frequent class represented the common homozygote. We summarised the genotype counts at each CNV using the *expected* posterior genotype calls (i.e. averaging over the posterior probabilities), from which we then calculated the minor allele frequency by,

$$f = \frac{2n_0 + n_1}{2},$$

where n_0 and n_1 are the expected genotype counts for the rare homozygote and heterozygote respectively. The distribution of minor allele frequency is shown in Supplementary Figure 21.

7.4 Power curves

We estimated power as follows. For a biallelic variant, the trend test statistic is known to approximately follow a χ^2_1 distribution with non-centrality parameter[154,155],

$$\eta = 2N\phi(1-\phi)\frac{(f_1 - f_0)^2}{f(1-f)},$$

where N is the total number of samples, ϕ the proportion of the samples that are cases, and f_1, f_0 and f are the (expected) frequencies of the risk allele in the cases, controls and the whole sample respectively. The odds ratio, θ, enters this expression via the allele frequencies, by relating f_1 to f_0,

$$f_1 = \frac{\theta f_0}{1 - f_0 + \theta f_0},$$

and the allele frequency in the whole sample is a function of that in the cases and controls,

$$f = \phi f_0 + (1-\phi) f_1,$$

which entails that in order to fully specify the distribution, we only need to specify the following four terms: the sample size (N), the proportion of cases (ϕ), the odds ratio (θ) and the risk allele frequency in the controls (f_0). To calculate the power given this result, we also need to specify a significance threshold (e.g. a p-value threshold).

We calculated the power at each CNV using the following assumptions:

- 2,000 cases and 3,000 controls ($N = 5000$ and $\phi = 0.4$),
- the allele frequency is the same as that observed in our sample,
- for two sets of p-value thresholds, 1×10^{-4} and 1×10^{-5},
- for a range of odds ratios, from 1 to 2.

We did this only for the non-duplicate, well-separated autosomal CNVs which were called with either 2 or 3 classes (consistent with it being biallelic). The CNVs were then split by MAF into two groups: those with MAF between 0.02 and 0.05 ('rare' - 426 CNVs) and those with MAF greater than 0.05 ('common' - 1854 CNVs). The power curves in Supplementary Figure 22 show the mean power across CNVs for each of these groups, for the range of odds ratios stated above.

8 Replication and validation of associated CNVs in WTCCC

8.1 Overview of approach

In order to obtain independent evidence to validate a CNV call within our WTCCC sample and/ or to seek independent evidence of association within a new case-control sample, we selected the least resource-intensive approach that was appropriate. Given that many of the CNVs of interest were well-tagged by one or more SNPs, in many cases it was possible to use SNP data as a proxy for the CNV. Because of the ease of SNP assays and the ready availability of genotyped or imputed SNP data in large case-control datasets, this was our preferred approach. In most cases it was possible to "look up" the relevant SNP association data for our own WTCCC1 disease sample (thus, providing validation of our CNV result) and also "look up" the results in the largest published/ publicly available dataset (meta-analysis or mega-analysis) for that disease (thus, providing a test of replication). If a CNV was well-tagged by a SNP but appropriate SNP data were not available for the disease of interest, SNP genotyping was undertaken within the laboratory of the relevant disease PI, as described below. Where there was no SNP tag available, breakpoint or direct quantitative CNV assays were designed, as described below. Such assays were undertaken within the laboratory of the relevant disease PI, with input from the central CNV group at Sanger. Replication results and an indication of the method used can be found in Supplementary Table 13.

8.2 Laboratory validation/ replication of CNVs for RA

CNVR3041.1 breakpoint determination

A PCR assay was generated using primers 3041_25_F and 3041_2627_R (see Supplementary Table 12) which flank the putative deletion region, the approximately 850bp product was cleaned and sequenced in both directions using the same primers and compared to the reference sequence to determine the breakpoint. The breakpoint was consistent in 4 samples (NA12156, NA06985, WTCCC134008, WTCCC133972), it mapped to a 2160bp deletion located at Chr6:113807650-113809810 of NCBI build 36.1, there was a 3 base (GAC) microhomology around the breakpoint site. The sequence surrounding the site is as follows: TGCATCTCTGATCCATTATTGCTA/GAC/ TTGTACTTTGTTTGCCCTTTATCTAAAA This work was undertaken at Sanger.

RA samples: First Sequenom iPlex assay for selected CNVs:

CNVR116.1 and CNVR1859.1 were tested in the RA cohort lab, University of Manchester on 3425 cases and 2758 controls. The CNVR116.1 assay did not use a good tagging SNP (r^2 = 0.55). Standard Sequenom iPlex protocol was applied, sourcing primers and probes from Metabion (sequences in Supplementary Table 12), PCR (HotStarTaq) reagents from Qiagen, dNTPs from Bioline and all other reagents and SpectroCHIPs from Sequenom.

RA samples: Second Sequenom custom assay for selected CNVs:

Four tagging SNPs (see Supplementary Table 12) with a high LD were established but these fell inside CNVR116.1 therefore homologous ratio assays were designed to amplify the *RHCE* and *RHD* gene simultaneously with the same PCR primers. The results are based on the assumption the SNP is not a true SNP but a fixed single base difference between the two genes, and this base is probed with the single base extension primer. In the case of rs28553519, the extension product from *RHCE* will be an A and a C from *RHD*. If *RHD* is not deleted at all, A:C should be 1:1. If there is a heterozygote deletion A:C should be 2:1. If there is a homozygote deletion A:C should be 1:0.

The breakpoint established by sequencing for CNVR3041.1 was used to generate a breakpoint Sequenom assay. In a deletion breakpoint assay three PCR primers are used, these are placed either side of the breakpoint and can generate a product for either wild-type or deleted genotypes, the probe is then positioned directly adjacent to the breakpoint so that the extension base is breakpoint specific. As assays for both ends of the breakpoint are designed for, in the case of a deletion the forward primer of the 5' assay is used as the forward primer in the 3' assay and vice versa for the other end of the breakpoint. This eliminates any requirement for the off line design of the third PCR primer, simplifying design. In the case of the 5' assay, extension primers extended by A are from the non deletion and those from T are from the deletion. If CNVR3041.1 is not deleted the ratio of T:A will be 0:1, 1:1 if there is a heterozygote deletion and 1:0 in the case of a homozygote deletion.

Both CNVR116.1 and CNVR3041.1 Sequenom assays could be run in the same plex using standard iPlex

sequenom methodology. The plex was carried out in the RA cohort lab on 3341 cases and 2517 controls (selected to ensure no bias from blood donor effects at RHD locus) using the same reagent sources as the first RA Sequenom plex except for primers sourced from IDT, (sequences in Supplementary Table 12).

8.3 Laboratory validation/ replication of CNVs for BC

CNVR8164.1 genotyping by PCR and Illumina 670K genotyping array

Replication data for testing the CNVR8164.1 association was available using 5653 Controls from the Wellcome Trust Case Control Consortium Two (WTCCC2) study and 3973 Breast Cancer samples from the Familial Breast Cancer Study. The Breast Cancer samples were typed on the Illumina 670K genotyping array and the WTCCC2 Controls were typed on the Illumina 1M array. CNVR8164.1 is targeted by 20 probes (19 SNPs, 1 monomorphic) on the 670K array and 26 probes on the 1M array (24 SNPs, 2 monomorphic).

A subset of samples were excluded from the analysis due to QC, resulting in 3660 Breast Cancer samples, 1689 of which were common to the WTCCC CNV study, and 5186 Controls of which 2037 samples were common to WTCCC CNV study samples.

CNV calling was performed using a novel hierarchical mixture model designed for Illumina data. The CNV calling results for CNVR8164.1 revealed three classes for both Cases and Controls. The call rate was 99.69% for the Breast Cancer samples and 99.36% for the Controls. The minor allele frequency for the Breast Cancer samples was estimated as 9.4% and 8.0% for the Controls. The concordance of calls for samples in both WTCCC and Breast Cancer was 99.58% while for Controls in both WTCCC CNV study and WTCCC2 was 99.16%.

An association test was performed using the Cochran-Armitage trend test. Using all the Illumina data samples the resulting p-value was 0.0007 (Odds ratio=1.19). However, when we tested for association using Illumina data on samples in common with the WTCCC CNV study the p-value was 0.0027 (Odds ratio=1.27). Considering only samples which were not in WTCCC and performing the association test gave the resulting p-value of 0.1204 (Odds ratio=1.12).

Independently, CNVR8164.1 was genotyped by a Taqman Real-Time PCR Assay in 1139 breast cancer cases and 870 controls. The assay was setup as a duplex reaction using RNase P as a reference sequence. The target assay was designed to the *APOBEC3B* Exon 2 sequence region, within CNVR8164.1, using *Applied Biosystems Primer Express* software (sequences in Supplementary Table 12). The oligonucleotide sequences were checked for absence of known SNP positions and for specificity, particularly with respect to *APOBEC3* paralogous gene homology. An endogenous control VIC-signal was obtained for all samples using the *Applied Biosystems* RNase P Endogenous Control Reagent Kit, primer limited, PN 4316844. PCR was performed in a 10μl reaction volume containing TaqMan Gene Expression Master Mix, 900nM Primers and 250nM probe for the target assay, RNase P reagent (supplied at x20 conc.) and 5ng (native) DNA, using standard Taqman quantification cycling parameters. Four replicate reactions were analysed for each sample, in a 384-well plate and all plates included three calibrator samples of copy number 2.

Calibrator ΔCt was calculated as the mean of the 12 values obtained and subsequently subtracted from all mean test sample ΔCts to obtain $\Delta\Delta$Ct. $2^{-\Delta\Delta Ct}$ provided the relative level of template present (RQ), which was multiplied by 2 to obtain sample copy number. For samples demonstrating presence of homozygous deletion allele, the level of endogenous control signal was confirmed as being typical for a functioning assay, to establish reaction and sample viability. An accuracy range of CN=1 or 2 $\pm$20% was considered acceptable, mean values falling outside this range were rejected. An association test was performed using the Cochran-Armitage trend test which gave a p-value of 0.2988. There were 827 breast cancer samples included in both replication experiments, Illumina and Taqman, and the concordance of calls between the two independent assays undertaken on these samples was 99.15%.

8.4 Laboratory validation/ replication of CNVs for T1D

CNVR7113.6

A Taqman genotyping assay was designed to assay SNP rs17426195, with the FAM probe calling the G allele and the VIC probe calling the A allele (Sequences in Supplementary Table 12). The assay was designed using the "Assays-by-Design" Service from Applied Biosystems and was run on the ABI 7900HT using ABI mastermix and standard protocol cycling conditions (95°C for 10mins, 40 cycles of 92°C for 15 seconds and 60°C for 1 minute, 10°C hold).

8.5 Laboratory validation/ replication of CNVs for CD

CNVR7113.6

A Sequenom assay was designed to assay SNP rs17426195 in 8,200 cases and 10,100 controls (Sequences in Supplementary Table 12). The Standard Sequenom iPlex protocol was applied, sourcing primers and probes from IDT, HotstarTaq DNA polymerase from Qiagen, dNTPs from Abgene and all other reagents and SpectroCHIPs from Sequenom.

9 Other analyses

9.1 Geographical stratification

An analysis was carried out to assess evidence of geographical stratification. At each CNV the CNV calls were regressed upon 13 binary covariates coding for the geographic region of origin of each sample. (Note that in WTCCC1 there were only 12 geographical regions used in analysis. In the current study, we use those same 12 regions plus one additional region for a small number of samples that came from Northern Ireland). The gender of each sample and binary covariates coding for WTCCC cohort were also included as baseline covariates. This results in a 12 degree-of-freedom test. CNVs included in this analysis were filtered on the basis of the clustering quality score combined with manual inspection of the most significant associations. Supplementary Figure 24 shows the distribution of -log10(p) along the 22 autosomal chromosomes. A significant effect was found in the HLA region as expected but no broad scale effects were uncovered.

9.2 Polymorphism analysis among single-class CNVs

For most of the CNVs on the array it was not possible to assign robust diploid copy number classes. This resulted in a final estimate of only one class at these CNVs. In order to ascertain how many of these CNVs are truly polymorphic in the WTCCC samples we looked at the variation of the normalized signal intensity between duplicate pairs. At polymorphic CNVs we should see much less variation in repeated measurements for duplicate samples than we should between non-duplicate samples. There will be some noise (eg. due to differences in DNA concentration or quality) but it should give a relative measure of the evidence for polymorphism at different CNVs. In order to be robust to outliers we used median absolute deviation (MAD) as the measure of spread:

Let D_i^j denote the j^{th} individual in the i^{th} duplicate pair and define

$$MAD_d = \frac{1}{2}\text{median}_i \left(\mid D_i^1 - D_i^2 \mid \right)$$

Denote the k^{th} sample by S_k and define

$$MAD_n = \text{median}_k \left(\mid S_k - M_s \mid \right)$$

where $M_s = \text{median}(S_k)$. Then we use

$$\frac{MAD_d}{MAD_n}$$

as the measure of the evidence for polymorphism at a CNV. The distribution of this statistic at CNVs called with 1 class is bimodal (Supplementary Figure 26). Also, the distribution of the MAD statistic at multi-class CNVs is much lower amongst the multi-class CNVs than amongst one-class and lies almost entirely under the left mode of the distribution (Supplementary Figures 26 and 25). That is, at multi class CNVs there is less variation between duplicate pairs than between non duplicate samples. Furthermore, at CNVs with higher numbers of classes we see that this effect increases.

In order to get estimates of the proportion of truly polymorphic CNVs amongst those called with only one class a threshold of 0.75 was used (Supplementary Figure 26). Using this threshold we find that $\sim 3624(57\%)$ of the CNVs assigned only 1 class show evidence for underlying polymorphism. To further explore this effect CNVs were broken down according to their discovery properties. Amongst CNVs discovered by the GSV consortium the CNVs were partitioned into four categories: (a) those discovered in multiple CEU samples; (b) those discovered in exactly one CEU sample; (c) those CNVs not discovered in any CEU samples but discovered in more than one YRI sample; (d) those CNVs not discovered in any CEU samples but discovered in exactly one YRI sample. As the WTCCC collections are European we would expect discoveries in the CEU to most reliably show polymorphism in the WTCCC data and this is reflected in our results (Supplementary Table 15).

9.3 Coverage of common autosomal CNVs in this study

Conrad et al. (2009) [103] have estimated that the number of common autosomal CNVs (MAF$> 5\%$) segregating in a population of European ancestry is 3,797. In order to estimate the number of autosomal CNVs with MAF$> 5\%$ that we have been able to test robustly for association in this study, we counted 1,585 likely biallelic CNVs (those with 2 or 3 copy number classes and with HWE p value > 0.0005) as having MAF $> 5\%$, from among the 3,339 autosomal CNVs passing our quality control filters. If we include the 321 CNVs with 4 or more copy number classes, for which we cannot reliably estimate minor allele frequencies, but are likely to have appreciable frequencies of non-reference alleles, then we obtain a total of 1,906 common CNVs that we have been able to assess in this study. Using these lower and upper estimates of the number of common CNVs studied here, we have assayed between 42% (1,585/3,797) and 50%

(1,906/3,797) of the estimated number of common autosomal CNVs segregating in a population of European ancestry.

9.4 Characterization of complexity at TSPAN8 locus

The most significant T2D associated CNV is CNVR5583.1, located in the tetraspanin/transmembrane 4 family gene *TSPAN8* region. In this gene region a T2D association has previously been reported[149] and the most significantly associated SNP is rs7961581 (p = 1.1×10^{-9}). However, the signal at this CNV is complex (Supplementary Figure 31A). A careful analysis of this region showed that the signal can be separated between three partially independent CNV signals: a 2-component CNV subset with a very rare third component (Supplementary Figure 31B), a clear 3-component CNV subset (Supplementary Figure 31C), and a more complex CNV subset with 5 distinct copy-number classes (Supplementary Figure 31D). Our automated association analysis initially identified the T2D association signal originating from the 3-component CNV (Supplementary Figure 31C), located between exons 4 and 5 of *TSPAN8*. However, subsequent analysis showed a potentially stronger association signal for the complex multi-allelic CNV (Supplementary Figure 31D). The genotypes at these three CNVs are partially correlated (see Supplementary Figure 32).

To characterize this CNV region, we designed a series of PCR primers to assay the three regions independently and selected samples for validation representing a range of different copy numbers for the three CNV subsets. We also compared our genotype data to the genotypes of three SNPs within the region (rs11178648, rs11178649 and rs3763978) using dbSNP. The results of this analysis showed concordance between the sequenced PCR products and the copy number classes detected in the main experiment, with a deletion event consistently sequenced in samples lying within the lowest copy number class for the 3-component and complex CNV subsets. Also, the detected genotypes for the three SNPs (homozygous positive, heterozygous, or homozygous negative) were consistent within samples. The homozygous negative genotype is consistently found in the R1 haplotype and the homozygous positive genotype is consistently found in the R0 haplotype and interaction with the I0 and I1 haplotypes changes the clustering within the 5-component CNV subsets. The sequencing data also suggest that the 3-component CNV (Supplementary Figure 31C) captures an inverted translocation or duplication.

While additional experimental work is required to completely characterize this region, available data already enabled a more detailed analysis of the T2D association signal. For this analysis we refer to both haplotypes defined by the clearest 3-component CNV (Supplementary Figure 31C) as R (for Reference) and I (for Inverted). Our data show that the complex multi-allelic CNV (Supplementary Figure 31D) only splits the R haplotype. Similarly, the rare two component CNV (Supplementary Figure 31B) only splits the I haplotype. These observations show that the genotype data from these three CNVs only capture four distinct haplotypes. We refer to the haplotypes defined by the 2-component CNV (Supplementary Figure 31B) as I0 and I1. We denote as R0 and R1 the haplotypes identified by the complex multi-allelic CNV (Supplementary Figure 31D). In our samples the control minor allele frequencies of the R0, R1, I0 and I1 haplotypes were 43%, 19%, 35% and 3%, respectively.

Using these notations for the four haplotypes (R0, R1, I0 and I1) we performed a haplotype analysis of the association signal. There was clear heterogeneity in disease risk in each category (Supplementary Figure 33), consistent with the detected signal. A logistic analysis including R0, I1 and I2 as covariates suggests that R0 is the main risk haplotype, but a more complex pattern involving additional haplotypes cannot be excluded. In particular, our analysis suggests a potential protective effect of the I1 haplotype (estimated odds ratio 0.8) but the low minor allele frequency of this haplotype (3%) prevents us from obtaining convincing evidence of association ($p = 0.086$). We also computed the level of linkage disequilibrium with rs7961581, which is the best SNP available from the most recent T2D meta-analysis[149]. Levels of LD were low with all the haplotypes we defined ($r^2 = 0.17, 0, 0.095$ and 0.036 with I0, I1, R0 and R1 respectively). Taken together, these data indicate that additional analytical and experimental work will be required to better understand the basis of the T2D/*TSPAN8* association.

The signal for the 3-component CNV is perfectly tagged ($r^2 = 1$) by three SNPs; rs1798090, rs1798089 and rs1705261. Our replication results (Supplementary Table 13) are based on rs1798090.

9.5 Comparing groups of CNVs

We performed two analyses to assess whether there is aggregate evidence that CNVs of a certain type might

be enriched amongst disease susceptibility loci. In particular, we considered: (i) CNVs which delete all or part of exons, compared to deletions that do not; and (ii) CNVs which are well-tagged by SNPs, compared to those poorly tagged.

The approach we took was, for each CNV in one group, to find a matching CNV in the other group and test whether there is a significant difference in the distribution of BFs between them. The criteria for CNVs to match is that they have the same number of classes and MAFs that differ by no more than 1%. Matching was done with replacement, to allow for situations where not enough CNVs in one set can match every CNV in the other. Within these constraints, the matching was done randomly. To average over possible matchings, we repeated the analysis 1000 times and report the median p-value from the test.

A two-sided Wilcoxon signed-rank test[156] was used to assess the significance of differences between the BF distributions. We did this separately for BFs calculated with and without use of the expanded reference panel; the two versions gave broadly similar results so we report only the latter.

When performing such analyses it is important to control for possible artefacts in the data. Only CNVs with high quality calls that passed the QC for testing were used. In addition, we also removed all CNVs in the HLA and Immunoglobulin regions, and CNVs with 4 or more classes (which are not biallelic and thus do not allow calculation of MAF or the usual r^2). To check that the aggregate BF differences are not due to artefacts, we also performed the above tests using the BFs from the control-control comparison.

Exonic deletions.

After QC filtering as described above, 53 deletions in exonic regions remained, all of which could be matched to a non-exonic deletion. No significant effects were observed for any of the diseases (see Supplementary Table 16), although this may reflect a lack of power.

CNVs well-tagged by SNPs.

As many of the CNVs on the Agilent chip are well-tagged by SNPs it is interesting to ask whether those CNVs not well-tagged by SNPs have different properties with respect to disease associations. For the purpose of this analysis, we defined a CNV to be *well-tagged* if it had $r^2 > 0.8$ for at least one SNP from one of the three collections of SNPs defined in Section 7.1.

For this analysis, we further excluded CNVs with MAF less than 1% and those with HWE p-value less than 1×10^{-10}, both of which gave rise to artefactual BF differences (as evidenced by the control-control BFs; data not shown). This left 1,485 well-tagged CNVs and 796 that were not well-tagged. Of these, 794 pairs could be successfully matched. No significant effects were observed for any of the diseases (see Supplementary Table 16).

10 Glossary

arc	Arthritis and rheumatism council
BC	Breast cancer
BD	Bipolar disorder
BIC	Bayesian Information Criteria
CAD	Coronary artery disease
BLAT	BLAST-like alignment tool
BMI	Body mass index
CD	Crohn's disease
CEPH	Centre dEtude Polymorphisme Humaine
CEU	Samples from European population used in HapMap project
CGH	Comparative Genomic Hybridization
CHB	Samples from Han Chinese from Beijing used in HapMap project
CNV	Copy number variation (or variant)
CNVE	Copy number variation event
CNVR	Copy number variation region
DLRS	Derivative log ratio spread
ECACC	European Collection of Animal Cell Cultures
EM	Expectation-maximization
GRID	Genetic Resource Investigating Diabetes
GSV	Genomic Structural Variation
GWAS	Genome-wide Association Study
HapMap	Large-scale international haplotype mapping studies undertaken to provide resource for association studies. The different integer suffices (1, 2 or 3) signify the dataset used.
hME	Homogenous Mass Extend assay (Sequenom Inc, San Diego, USA)
HT	Hypertension
Ig	Immunoglobulin
iPLEX	Multiplex SNP genotyping assay (Sequenom Inc, San Diego, USA)
JDRF	Juvenile Diabetes Research Foundation
JPT	Samples from Japanese population in Tokyo used in HapMap project
LD	Linkage Disequilibrium
MAD	Median absolute deviation
MAP	Maximum a posteriori
MI	Myocardial infarction
MLPA	Multiplex Ligation-dependent Probe Amplification
MODY	Maturity onset diabetes of the young
NHSBT	National Health Service Blood Transfusion service
OGT	Oxford Gene Technology
OPCRIT	OPerational CRITeria checklist of psychotic symptoms
PCA	Principal Component Analysis
PI	Principal investigator
PNDM	Permanent neonatal diabetes
PVS	Probe variance scaling
RA	Rheumatoid arthritis
RR	Relative risk
SNP	Single nucleotide polymorphism
T1D	Type 1 diabetes
T2D	Type 2 diabetes
SNBTS	Scottish National Blood Transfusion Service
WBS	Welsh Blood Service
WT	Wellcome Trust
WTCCC	Wellcome Trust Case Control Consortium
WTCCC1	Wellcome Trust Case Control Consortium: SNP genome-wide association study (the first WTCCC study funded)
WTCCC2	Wellcome Trust Case Control Consortium: Genome-wide association study of a new set of disease and traits (a separate study from WTCCC1 and WTCCC)
YRI	Samples from Yoruban population from West African used in HapMap project

11 Acknowledgements

The principal funder of this project was the Wellcome Trust. Case collections were funded by: Arthritis Research Campaign, BDA Research, British Heart Foundation, British Hypertension Society, Canada Foundation of Innovation, Canadian Institutes of Health Research (CIHR), Cancer Research UK; Department of Pathology at Brigham and Womens Hospital , Diabetes UK; Department of Health (National Institute for Health Research Biomedical Research Centre Awards (Guy's & St Thomas' NHS Foundation Trust; King's College London; Cambridge University Hospitals NHS Foundation Trust; University of Cambridge School of Clinical Medicine; Central Manchester Foundation Trust; University of Manchester); Diabetes UK, European Commission (Framework 6); Genome Canada/Ontario Genomics Institute , Glaxo-SmithKline Research and Development, Juvenile Diabetes Research Foundation, Hospital for Sick Children Foundation , McLaughlin Centre for Molecular Medicine , Medical Research Council, National Association for Colitis and Crohns disease, Netherlands Organization for Scientific Research , Ontario Innovation Trust, Ontario Ministry of Research and Innovation , SHERT (The Scottish Hospitals Endowment Research Trust), St Bartholomews and The Royal London Charitable Foundation, UK Medical Research Council, UK NHS R&D; US Military ACQ Activity, US National Institutes of Health (NIH) and the Wellcome Trust. Statistical analyses were funded by the Wellcome Trust, the National Institutes of Health, and the Royal Society.

We acknowledge the many physicians, research fellows and research nurses who contributed to the various case collections, and the collection teams and senior management of the UK Blood Services responsible for the UK Blood Services Collection. For the 1958 Birth Cohort, venous blood collection was funded by the UK Medical Research Council and cell-line production, DNA extraction and processing by the Juvenile Diabetes Research Foundation and the Wellcome Trust. We recognize the contributions of: P. Shepherd (1958 Birth Cohort); those at OGT (particularly Graham Speight, Nicole Sparkes, Andrew Rogers, John Shovelton.) and Agilent (particularly Shane Giles, Sharoni Jacobs, Dione Bailey) for CNV assay optimization, data production and data delivery; A. Ardern-Jones, G. Attard, K. Bailey, C. Bardsley, J. Barwell, L. Baxter, R. Belk, J. Berg, N. Bradshaw, A. Brady, S. Brant, C. Brewer, G. Brice, G. Bromilow, C. Brooks, A. Bruce, B. Bulman, L. Burgess, J. Campbell, B. Castle, R. Cetnarskyj, C. Chapman, C. Chu, N. Coates, A. Collins, J. Cook, S. Coulson, G. Crawford, D. Cruger, C. Cummings, R. Davidson, L. Day, L. de Silva, B. Dell, C. Dolling, A. Donaldson, A. Donaldson, H. Dorkins, F. Douglas, S. Downing, S. Drummond, J. Dunlop, S. Durrell, D. Eccles, C. Eddy, M. Edwards, E. Edwards, J. Edwardson, R. Eeles, F. Elmslie, G. Evans, B. Gibbens, C. Giblin, S. Gibson, S. Goff, S. Goodman, D. Goudie, L. Greenhalgh, J. Greer, H. Gregory, R. Hardy, C. Hartigan, T. Heaton, C. Higgins, S. Hodgson, T. Homfray, D. Horrigan, C. Houghton, L. Hughes, V. Hunt, L. Irvine, L. Izatt, L. Jackson, C. Jacobs, S. James, M. James, L. Jeffers, I. Jobson, W. Jones, S. Kenwrick, C. Kightley, C. Kirk, L. Kirk, E. Kivuva, A. Kumar, F. Lalloo, N. Lambord, C. Langman, P. Leonard, S. Levene, S. Locker, P. Logan, M. Longmuir, A. Lucassen, V. Lyus, A. Magee, S. Mansour, D. McBride, E. McCann, V. McConnell, M. McEntagart, K. McDermot, L. McLeish, D. McLeod, L. Mercer, C. Mercer, Z. Miedzybrodzka, J. Miller, P. Morrison, J. Myring, J. Paterson, P. Pearson, G. Pichert, K. Platt, M. Porteous, C. Pottinger, S. Price, L. Protheroe, L. Protheroe, S. Pugh, N. Rahman, C. Riddick, V. Roffey-Johnson, M. Rogers, S. Rose, S. Rowe, A. Schofield, G. Scott, J. Scott, A. Searle, S. Shanley, S. Sharif, J. Shaw, J. Shea-Simonds, L. Side, J. Sillibourne, K. Simon, S. Simpson, S. Slater, K. Smith, L. Snadden, J. Soloway, Y. Stait, B. Stayner, M. Steel, C. Steel, H. Stewart, D. Stirling, M. Thomas, S. Thomas, S. Tomkins, H. Turner, E. Tyler, E. Wakeling, F. Waldrup, L. Walker, L. Walker, C. Watt, S. Watts, A. Webber, C. Whyte, J. Wiggins, E. Williams, L. Winchester (clinicians and counsellors from the Breast Cancer Susceptibility Collaboration (BCSC) who coordinated recruitment and collection of the BC samples); C. Fraser, G. Fraser, J. Heron, S. Hyde, A.Massey; F. Oyebode, M. Sinclair, A. Stern, N. Walker and S. Zammitt (recruitment and phenotypic assessment of BD cases); M. Yuille, B. Ollier and the UK DNA Banking Network and members of the BHF Family Heart Study Research Group (CAD case recruitment and DNA provision); S. Goldthorpe, D. Soars and J. Whittaker for CD collections; Members of UKRAG (UK Rheumatoid Arthritis Group) (RA case recruitment); J. Pembroke, M. Bruce, S. Colville-Stewart, K. Edwards, L. Gatherer, C. Gemmell, K. Gilmour, S. Hampson, S. Hood, J. Hunt, J. Hussein, J. Jamieson, J. Kent, D. Lloyd, K. MacFarlane, S. Mellow, A. Nixon, J. Pheby, D. Picton, F. Porteus, P. Whitworth, K. Witte, A. Zawadzka, C. Mein and the research nurses and the membership of the British Society for Paediatric Endocrinology and Diabetes, and David Dunger and the Department of Paediatrics, University of Cambridge (T1D case recruitment); M. Samp-

son, J.C. Levy, S. ORahilly, S. Howell, M. Murphy and A. Wilson (T2D case recruitment).

Essential informatics support was provided by the administration, systems, bioinformatics, data services and DNA teams of the JDRF/WT DIL; At the Sanger Institute we thanks: the Web System teams (particularly R. Pettitt); D. Holland and R. Vincent. T. Dibling, C. Hind, D. Simpkin, P. Ewels and D. Moore for genotyping assistance; Alagu Jayakumar, Simon Potter and Paul Weston for informatics support; Jackie Bryant, Thomas Dibling, Alicja Wilk, Stephen Gamble, Radhi Ravindrarajah, and Adam Whittaker for help with sample logistics; Yujun Zhang and Tomas Fitzgerald for advice on replication/ validation activities; Suzannah Bumpstead for help with sample handling.

We thank all members of major international research consortia or collaborations who have facilitated the work of this study, including the Genome Structural Variation (GSV) Consortium (including thanks to Deanne Church, Steve McCarroll, Peggy Eis, Todd Richmond, Michael Hogan [Nimblegen]) and the Global BP Gen Consortium, and the Wellcome Trust Case Control Consortium 2 for making available SNP data on UKBS individuals for our tagging analyses.

Personal support was provided by: British Heart Foundation (S.J.B., N.J.S., A.Do., M.C., M.B.); GlaxoSmithKline (S.W.S), SIM (G.B.); UK Medical Research Council (C.T., M.E.T.); US Military ACQ Activity, Era of Hope Award (D.D, D.H.); Royal Netherlands Academy of Arts and Sciences (D.P.), Vandervell Foundation (M.N.W.); and Wellcome Trust (A.P.M., C.M.L. E.Z., R.M.F).

12 Author contributions

12.1 The contributions of the authors are as follows:

WTCCC Management Committee: Peter Donnelly (Chair)[1,2], Nick Craddock[3], Panos Deloukas[4], Audrey Duncanson[5], Matthew E Hurles[4], Dominic P Kwiatkowski[1,4], Mark I McCarthy[1,6,7], Willem H Ouwehand[4,8,9], Miles Parkes[10], Nazneen Rahman[11], Nilesh J Samani[12,13], John A Todd[14].

WTCCC CNV Committee: Nick Craddock[3] (co-chair), Matthew E Hurles[4] (co-chair), Chris Barnes[4], Niall Cardin[2], Donald F Conrad[4], Peter Donnelly[1,2], Eleni Giannoulatou[2], Chris Holmes[2], Jonathan L Marchini[2], Richard D Pearson[1], Vincent Plagnol[14], Samuel Robson[4], Nilesh J Samani[12,13], Kathy Stirrups[4], Martin D Tobin[15], Damjan Vukcevic[1], Louise V Wain[15], Chris Yau[2].

WTCCC Fine-mapping and Resequencing Committee: Panos Deloukas[4] (co-chair), Peter Donnelly[1,2] (co-chair), Dominic P Kwiatkowski[1,4] (co-chair), Mark I McCarthy[1,6,7] (co-chair), Inês Barroso[4], Matthew A Brown[16,17], Gil McVean[2], Simon Myers[2], Willem H Ouwehand[4,8,9], Aarno Palotie[4], Miles Parkes[10], Nazneen Rahman[11], Nilesh J Samani[12,13], John A Todd[14], Jane Worthington[18], Eleftheria Zeggini[1,4].

Autoimmune Thyroid Disease: Oliver J Brand[19], Jayne A Franklyn[19,20], Matthew J Simmonds[19], Stephen CL Gough[19,20].

Ankylosing Spondylitis: David M Evans[21], Millicent A Stone[22,23], B Paul Wordsworth[16], Matthew A Brown[16,17].

Breast Cancer: Jaswinder Bull[11], Darshna Dudakia[11], Bernadette Ebbs[11], Diana Eccles[24], Anna Elliot[11], Gareth Evans[25], Polly Gibbs[11], Anita Hall[11], Sarah Hines[11], Debbie Hughes[11], David Pernet[11], Anthony Renwick[11], Richard Scott[11], Sheila Seal[11], Katarina Spanova[11], Clare Turnbull[11], Margaret Warren-Perry[11], Michael R Stratton[4,11], Nazneen Rahman[11].

Bipolar Disorder: Gerome Breen[26,27], Sian Caesar[28], Anne Farmer[27], I Nicol Ferrier[29], Liz Forty[3], Katherine Gordon-Smith[3,28], Elaine Green[3], Detelina Grozeva[3], Ian R Jones[3], Lisa A Jones[28], George Kirov[3], Peter McGuffin[27], Michael C O'Donovan[3], Michael J Owen[3], Ellie Russell[3], David St Clair[26], Allan H Young[29,30], Nick Craddock[3].

Coronary Artery Disease: Stephen G Ball[31], Anthony J Balmforth[31], Peter S Braund[12], Paul R Burton[15], Panos Deloukas[4], Suzanne Rafelt[12], John R Thompson[15], Alistair S Hall[31], Nilesh J Samani[12,13].

Crohns Disease: Tariq Ahmad[32], Jeffrey C Barrett[4], Katarzyna Blaszczyk[33], Francesca Bredin[10], Hazel E Drummond[34], Cathryn Edwards[35], Alistair Forbes[36], Mahim Jain[1], Derek P Jewell[37], Charlie Lees[34], James Lee[10], John Mansfield[38], Dunecan C O Massey[10], Alex Mentzer[39], Craig Mowat[40], William Newman[25], Elaine R Nimmo[34], Anne Phillips[40], Natalie J Prescott[33], Jeremy D Sanderson[39], Jack Satsangi[34], Christopher G Mathew[33], Miles Parkes[10].

Hypertension: Morris J Brown[41], John MC Connell[42], Anna F Dominiczak[42], Martin Farrall[43], Philip Howard[44], Toby Johnson[44], G Mark Lathrop[45], Kate L Lee[44], Abiodun Onipinla[44], Nilesh J Samani[12,13], Sue Shaw-Hawkins[44], John Webster[46], Patricia B Munroe[44], Mark J Caulfield[44].

Multiple Sclerosis: Alastair Compston[47], Stephen J Sawcer[47].

Rheumatoid Arthritis: Anne Barton[18], John Bowes[18], Ian N Bruce[18], Paul Emery[48], Steve Eyre[18], Edward Flynn[18], Paul Gilbert[18], Pile Harrison[49], Anne Hinks[18], Lynne Hocking[50], John D Isaacs[51], Paul Martin[18], Ann E Morgan[52], David M Reid[50], Deborah PM Symmons[18], Sophia Steer[53], Wendy Thomson[18], Anthony G Wilson[54], B Paul Wordsworth[16], Jane Worthington[18].

Type 1 Diabetes: Oliver S Burren[14], Jason D Cooper[14], Kate Downes[14], Matt Hardy[14], Joanna MM Howson[14], Meeta Maisuria-Armer[14], Nigel R Ovington[14], Vincent Plagnol[14], Helen Schuilenburg[14], Debbie J Smyth[14], Helen E Stevens[14], Neil M Walker[14], Chris Wallace[14], Matthew Woodburn[14], John A Todd[14].

Type 2 Diabetes: Amanda J Bennett[6], Rachel M Freathy[55], Chris J Groves[6], Neelam Hassanali[6], Graham A Hitman[56], Hana Lango-Allen[55], Cecilia M Lindgren[1,6], Andrew P Morris[1], Kirstie Parnell[55], John

RB Perry[55], Inga Prokopenko[1,6], Nigel W Rayner[1,6], Neil Robertson[1,6], Beverley M Shields[55], Mary E Travers[6], Mark Walker[57], Michael N Weedon[55], Eleftheria Zeggini[1,4], Andrew T Hattersley[55,58], Timothy M Frayling[55], Mark I McCarthy[1,6,7].

1958 Birth Cohort Controls: Wendy L McArdle[59], Susan M Ring[59], David P Strachan[60].

UK Blood Services Controls: Anthony Attwood[4,8,9], Jennifer D Jolley[8,9], Jennifer G Sambrook[8,9], Jonathan Stephens[8,9], Nicholas A Watkins[8,9], Willem H Ouwehand[4,8,9].

Sample processing, genotyping and sequencing: Hazel Arbury[4], Sanjeev Bhaskar[4], John Burton[4], Chris M Clee[4], Alison J Coffey[4], Andrew Dunham[4], Sarah Edkins[4], Emma Gray[4], Rhian Gwilliam[4], Husam Hebaishi[4], Eleanor Howard[4], Sarah Hunt[4], Cordelia F Langford[4], Kirsten E McLay[4], Michael L Mimmack[4], Michael A Quail[4], Elilan Somaskantharajah[4], Aarno Palotie[4], Panos Deloukas[4].

CNV Analysis: Jan Aerts[4], Chris Barnes[4], Niall Cardin[2], Donald F Conrad[4], Nick Craddock[3], Eleni Giannoulatou[2], Naomi Hammond[4], Chris Holmes[2], Kevin Lewis[4], Jonathan L Marchini[2], Richard D Pearson[1], Vincent Plagnol[14], Samuel Robson[4], Kathy Stirrups[4], Martin D Tobin[15], Damjan Vukcevic[1], Louise V Wain[15], Chris Yau[2], Peter Donnelly[1,2], Matthew E Hurles[4].

Fine-mapping and sequence analysis: Adam Auton[2], Jake Byrnes[1], Sarah Hunt[4], Julian Maller[1], Andrew P Morris[1], Simon Myers[2], Gil McVean[2], Kimmo Palin[4], Carol E Scott[4], Zhan Su[2], Peter Donnelly[1,2].

Genome Structural Variation Consortium: Donald F Conrad[4], Dalila Pinto[61], Richard Redon[4,62], Lars Feuk[61,63], Omer Gokumen[64], Jan Aerts[4], T Daniel Andrews[4], Chris Barnes[4], Tomas Fitzgerald[4], Ifejinelo Onyiah[4], Armand Valsesia[4], Chris Tyler-Smith[4], Nigel P Carter[4], Charles Lee[64], Stephen W Scherer[61,65], Matthew E Hurles[4].

CNV Writing Group: Nick Craddock[3], Matthew E Hurles[4], Jonathan L Marchini[2], Richard D Pearson[1], Vincent Plagnol[14], Nilesh J Samani[12,13], Peter Donnelly[1,2].

WTCCC Publications Committee: Nick Craddock[3], Peter Donnelly[1,2], Willem H Ouwehand[4,8,9], Nilesh J Samani[12,13], Mark I McCarthy[1,6,7] (Chair).

WTCCC Principal Investigators: Matthew A Brown[16,17], Paul R Burton[15], Mark J Caulfield[44], Alastair Compston[47], Nick Craddock[3], Panos Deloukas[4], Martin Farrall[43], Stephen CL Gough[19,20], Alistair S Hall[31], Andrew T Hattersley[55,58], Adrian VS Hill[1], Matthew E Hurles[4], Dominic P Kwiatkowski[1,4], Mark I McCarthy[1,6,7], Christopher G Mathew[33], Willem H Ouwehand[4,8,9], Miles Parkes[10], Marcus Pembrey[66], Nazneen Rahman[11], Nilesh J Samani[12,13], Jack Satsangi[34], Michael R Stratton[4,11], John A Todd[14], Jane Worthington[18], Peter Donnelly[1,2].

12.2 Affiliation List For The Author Contribution Listing

[1] The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.

[2] Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.

[3] MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK.

[4] The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA UK.

[5] The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.

[6] Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK.

[7] Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, OX3 7LJ, UK.

[8] Department of Haematology, University of Cambridge, Long Road, Cambridge, CB2 0PT, UK.

[9] National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge CB2 0PT, UK.

[10] IBD Genetics Research Group, Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK.

[11] Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, UK.

[12] Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester LE3 9QP, UK.

[13] Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, LE3

9QP, UK.

[14] Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK.

[15] Departments of Health Sciences and Genetics, University of Leicester, 217 Adrian Building, University Road, Leicester, LE1 7RH, UK.

[16] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Nuffield Orthopaedic Centre, University of Oxford, Windmill Road, Headington, Oxford, OX3 7LD, UK.

[17] Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Ipswich Road, Woolloongabba, Brisbane, Queensland, 4102, Australia.

[18] arc Epidemiology Unit, Stopford Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK.

[19] Centre for Endocrinology, Diabetes and Metabolism, Institute of Biomedical Research, University of Birmingham, Birmingham, B15 2TT, UK.

[20] University Hospital Birmingham NHS Foundation Trust, Birmingham, B15 2TT, UK.

[21] MRC Centre for Causal Analyses in Translational Epidemiology, Department of Social Medicine, University of Bristol, Bristol, BS8 2BN, UK.

[22] University of Toronto, St. Michael's Hospital, 30 Bond Street, Toronto, Ontario M5B 1W8, Canada.

[23] University of Bath, Claverdon, Norwood House, Room 5.11a Bath Somerset BA2 7AY UK.

[24] Academic Unit of Genetic Medicine, University of Southampton, Southampton, UK.

[25] Department of Medical Genetics, Manchester Academic Health Science Centre (MAHSC), University of Manchester, Manchester M13 0JH, UK

[26] University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK.

[27] SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK.

[28] Department of Psychiatry, University of Birmingham, National Centre for Mental Health, 25 Vincent Drive, Birmingham, B15 2FG, UK.

[29] School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne, NE1 4LP, UK.

[30] UBC Institute of Mental Health, 430-5950 University Boulevard Vancouver, British Columbia, V6T 1Z3, Canada.

[31] Multidisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds, LS2 9JT, UK.

[32] Genetics of Complex Traits, Peninsula College of Medicine and Dentistry University of Exeter, EX1 2LU, UK.

[33] Department of Medical and Molecular Genetics, Kings College London School of Medicine, 8th Floor Guys Tower, Guys Hospital, London, SE1 9RT, UK.

[34] Gastrointestinal Unit, Division of Medical Sciences, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK.

[35] Endoscopy Regional Training Unit, Torbay Hospital, Torbay TQ2 7AA, UK

[36] Institute for Digestive Diseases, University College London Hospitals Trust, London NW1 2BU, UK.

[37] Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford, OX2 6HE, UK.

[38] Department of Gastroenterology & Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK.

[39] Division of Nutritional Sciences, King's College London School of Biomedical and Health Sciences, London SE1 9NH, UK.

[40] Department of General Internal Medicine, Ninewells Hospital and Medical School, Ninewells Avenue, Dundee DD1 9SY

[41] Clinical Pharmacology Unit, University of Cambridge, Addenbrookes Hospital, Hills Road, Cambridge CB2 2QQ, UK.

[42] BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow, G12 8TA, UK.

[43] Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.

[44] Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK.

[45] Centre National de Genotypage, 2, Rue Gaston Cremieux, Evry, Paris 91057, France.

[46] Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK.

[47] Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road,

Cambridge, CB2 2QQ, UK.

[48] Academic Unit of Musculoskeletal Disease, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK.

[49] University of Oxford, Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford, OX3 7LD, UK.

[50] Bone Research Group, Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen, AB25 2ZD, UK.

[51] Institute of Cellular Medicine, Musculoskeletal Research Group, 4th Floor, Catherine Cookson Building, The Medical School, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK.

[52] NIHR-Leeds Musculoskeletal Biomedical Research Unit, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK.

[53] Clinical and Academic Rheumatology, Kings College Hosptal National Health Service Foundation Trust, Denmark Hill, London SE5 9RS, UK.

[54] School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield, S10 2JF, UK.

[55] Genetics of Complex Traits, Peninsula College of Medicine and Dentistry, University of Exeter, Magdalen Road, Exeter, EX1 2LU, UK.

[56] Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London, E1 1BB, UK.

[57] Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK.

[58] Genetics of Diabetes, Peninsula College of Medicine and Dentistry, University of Exeter, Barrack Road, Exeter, EX2 5DW, UK.

[59] ALSPAC Laboratory, Department of Social Medicine, University of Bristol, BS8 2BN, UK.

[60] Division of Community Health Sciences, St George's, University of London, London SW17 0RE, UK.

[61] The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, MaRS Centre- East Tower, 101 College St, Room 14-701, Toronto, Ontario M5G 1L7, Canada

[62] INSERM UMR915, LInstitut du Thorax, Nantes, 44035, France

[63] Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala 75185, Sweden

[64] Department of Pathology, Brigham and Womens Hospital and Harvard Medical School, Boston MA02115, USA

[65] Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada

[66] Clinical and Molecular Genetics Unit, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK.

13 Figures

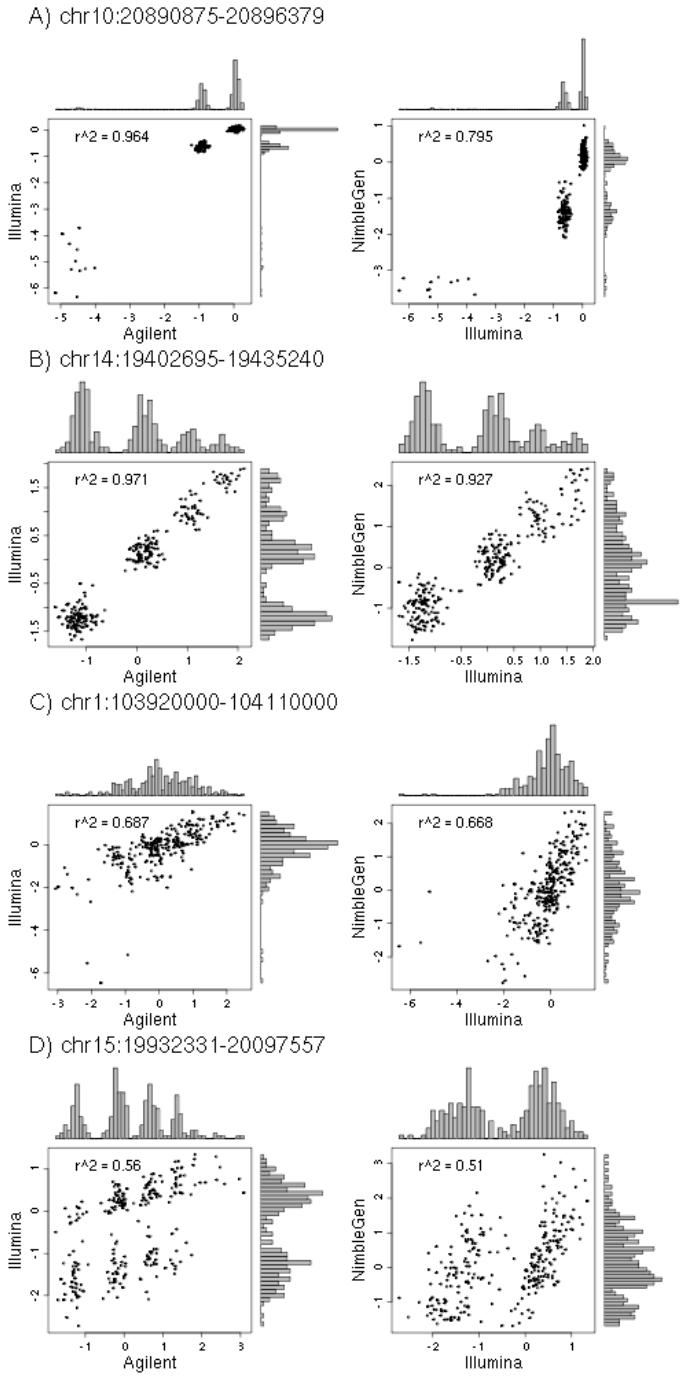


Figure 1: Bivariate plots comparing CNV signal data for the three genotyping arrays across four distinct CNV regions. Bivariate plots compare Agilent/Illumina data (left) and NimbleGen/Illumina data (right). The four CNVs were chosen to represent the variability in data quality in the pilot data: (A) shows a well clustered deletion with high concordance between the three platforms, (B) shows a more complex multi-allelic CNV with high concordance between platforms, (C) shows a CNV with insufficient data quality for clustering for all platforms and (D) shows a CNV where the Agilent platform has identified the complex multi-modal nature of this region, whilst Illumina and NimbleGen appear much more noisy.

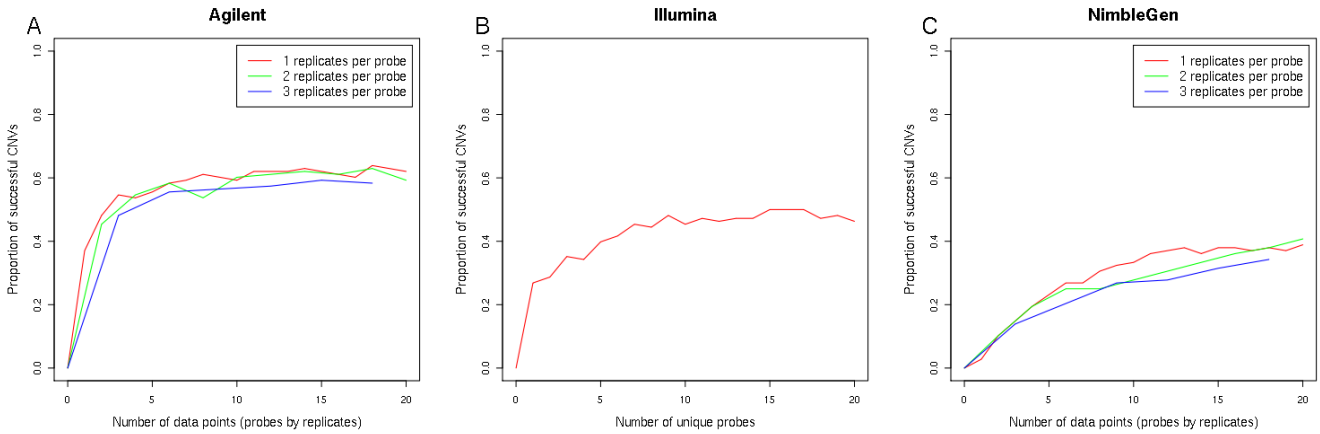


Figure 2: Effect of the number of probes per CNV on the overall number of clusterable CNVs for the Agilent (A), Illumina (B) and NimbleGen (C) arrays. The availability of probe replicates for the Agilent and NimbleGen array gave us the opportunity to explore the effect of probe replication (A and C).

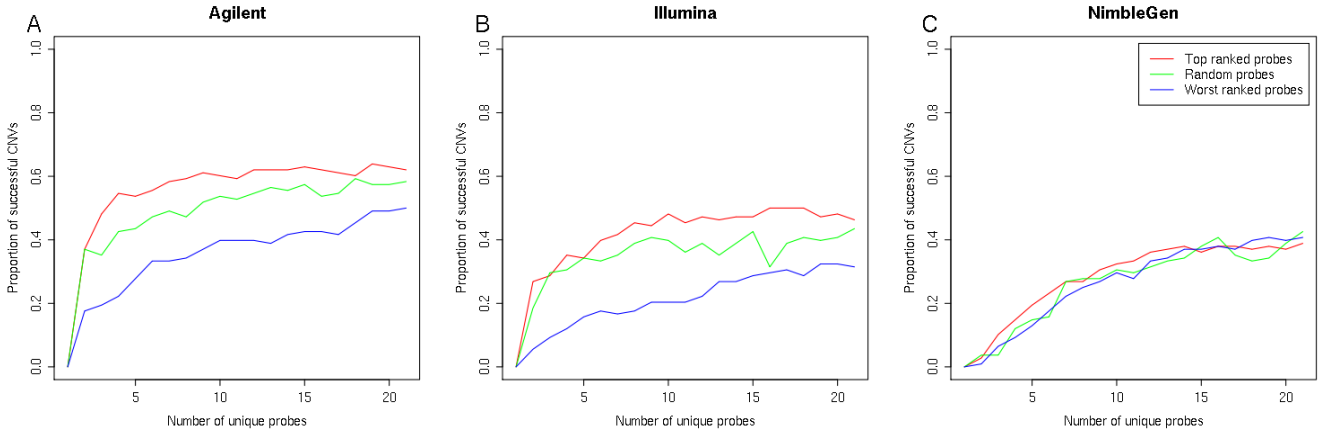


Figure 3: Effect of using platform-specific probe quality metrics to aid in the array design process. Probes were ranked based on metrics provided by each company and, for each platform, the overall proportion of well clustered CNVs is shown as a function of the number of probes used in the first principal component summary. The results are shown following selection from the top-ranking probes (red), the worst ranking probes (blue) and a random subset of the probes (green), and in each case, a single replicate of each probe was used. For all platforms, and particularly for the Agilent and Illumina data, using the probe-quality metric to influence selection had a significant impact on the quality of data and our ability to successfully genotype CNVs. Using a random subset of probes was typically more comparable to using the best ranked probes, suggesting that the probes available for each platform are generally of a high quality.

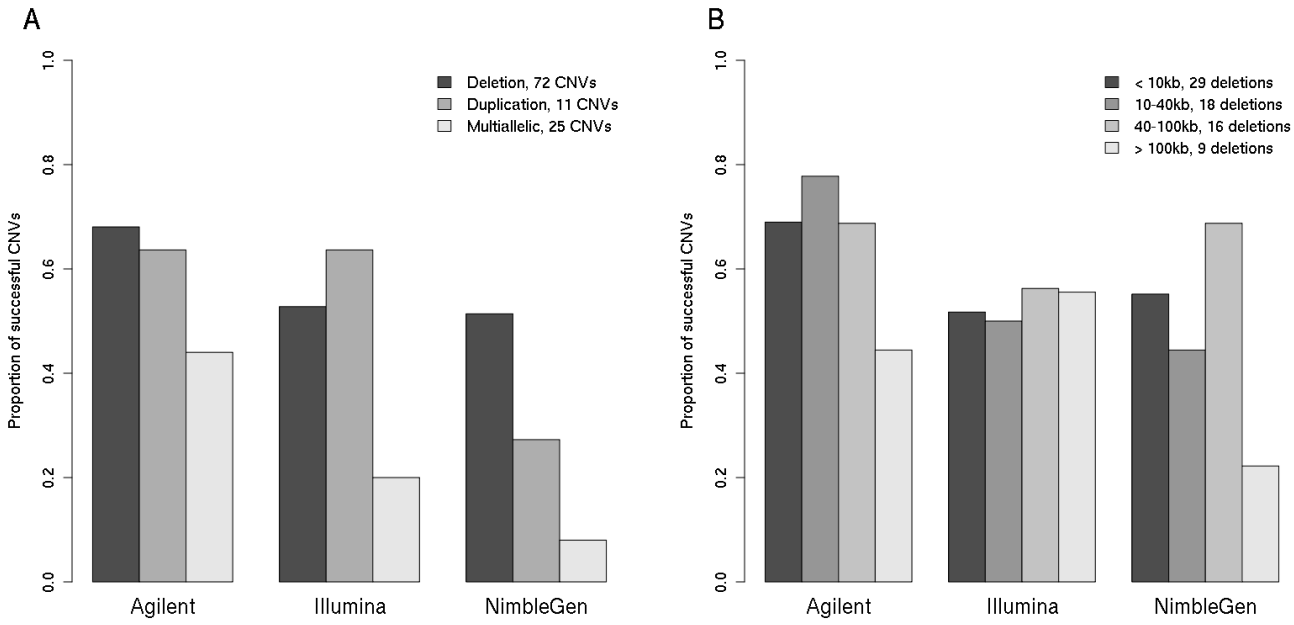


Figure 4: Comparison of the genotyping success rate for the three arrays separated by CNV class and CNV size.

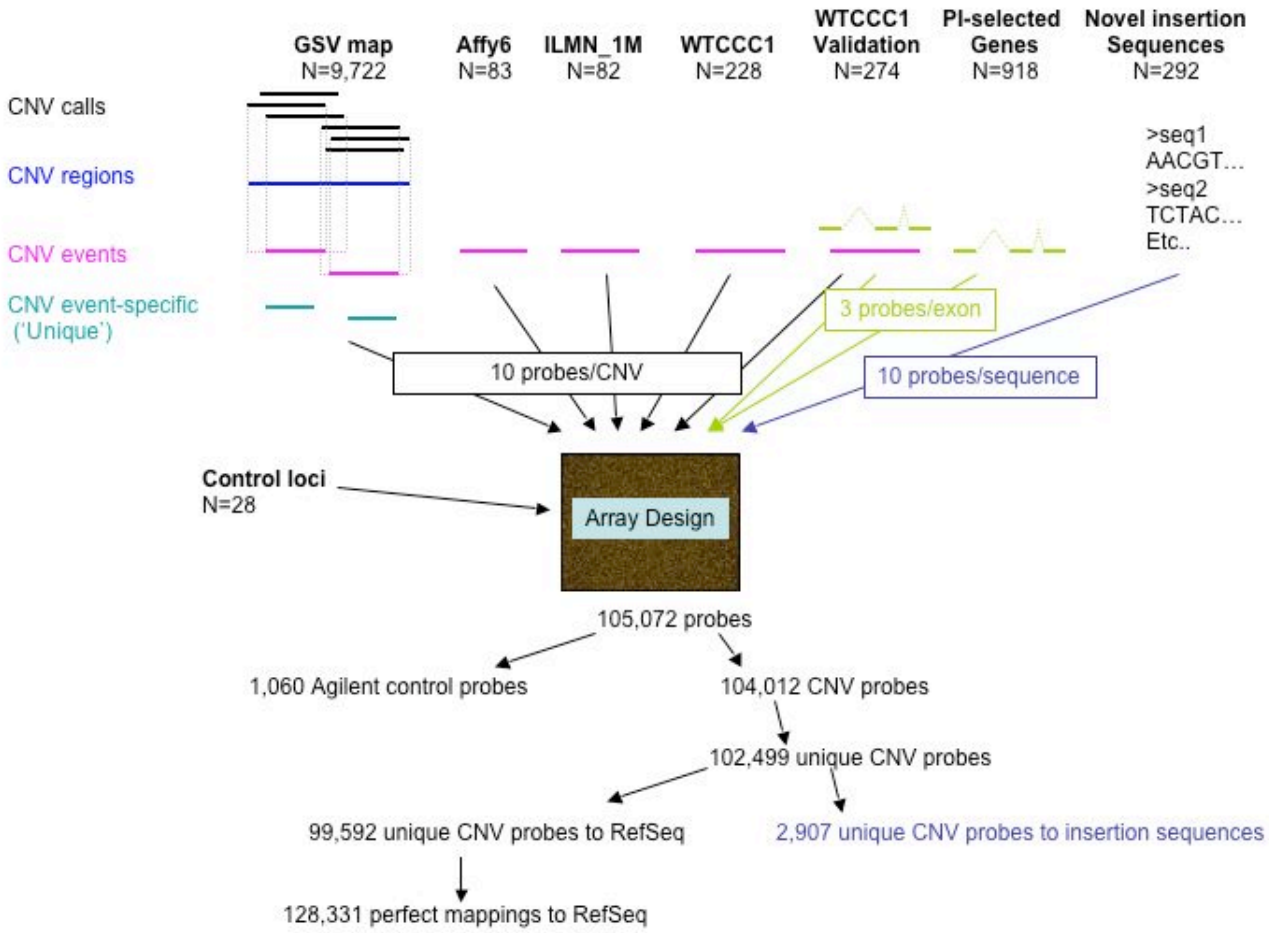


Figure 5: Summary of the content of the array. Schematic representation of the probe content of the designed array and the source of the putative copy number variation targeted by the design .

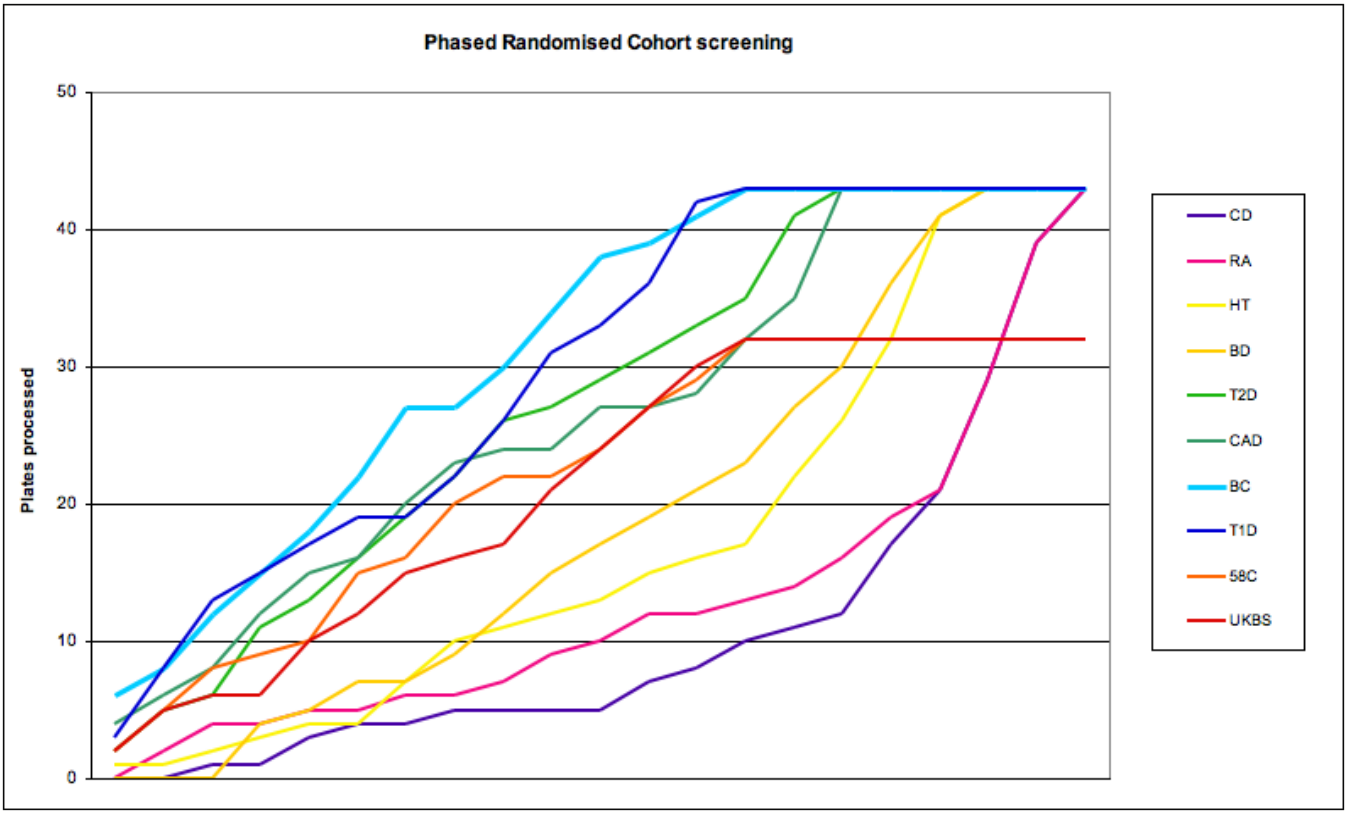


Figure 6: Phased Randomised Cohort screening. Schematic representation of time course of processing plates for the sample cohorts in WTCCC. Time from the start of the experiment is shown on the x-axis. Number of completed processed plates is shown on the y-axis. Phasing was used in order to provide completed datasets for controls and some case sets before the end of the experimental period in order to facilitate piloting of data analysis pipelines.

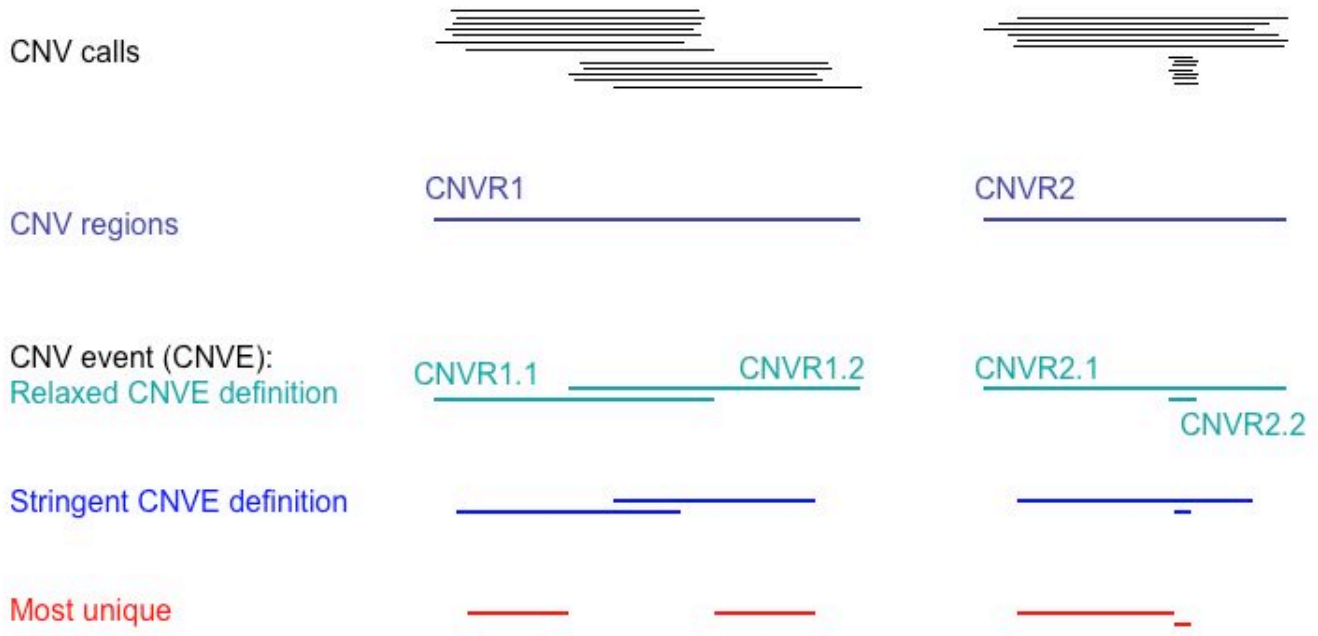


Figure 7: Schematic plot showing how most unique regions of each CNV from the GSV discovery project were defined.

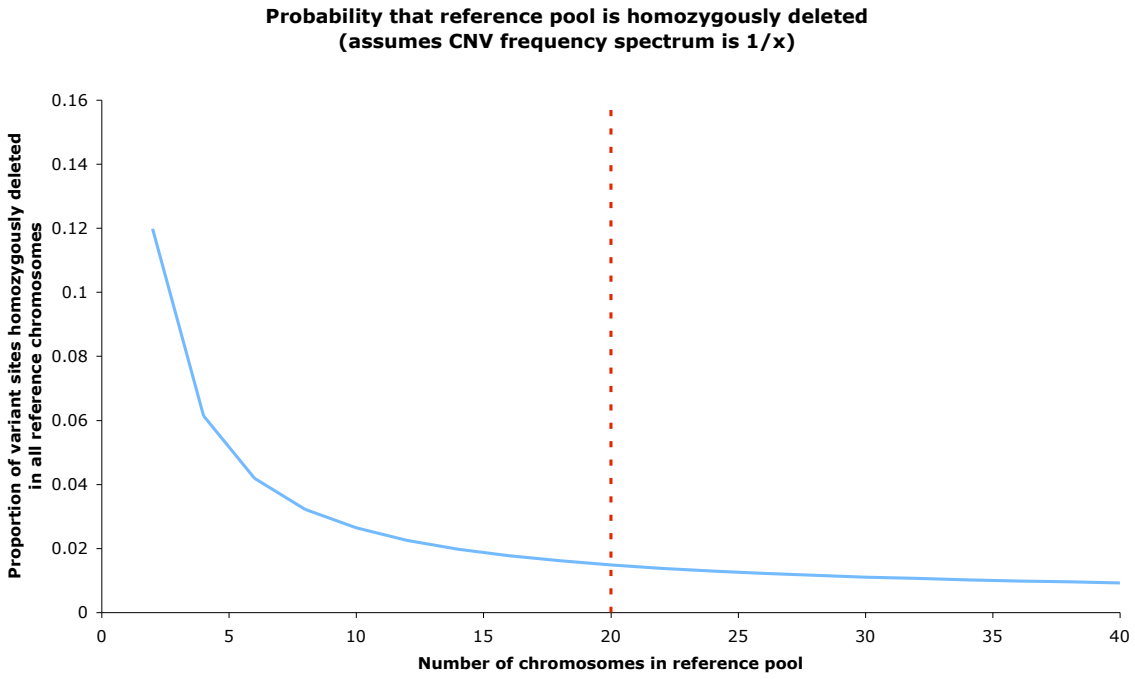


Figure 8: Proportion of CNVs at which the reference pool of chromosomes all have the deleted allele. At a given frequency (f) of the deletion allele in the population, the probability that the deleted allele is present in all N chromosomes in the reference pool is f^N . Here the frequency spectrum of the deleted allele is assumed to be that predicted by the infinite sites model (the 1/x model) applied to the discovery sample of 40 CEU chromosomes. Thus the proportion of loci at which all reference chromosomes contain the deleted allele can be estimated for each value of N. The red line shows the number of chromosome present in the reference pool used in this experiment.

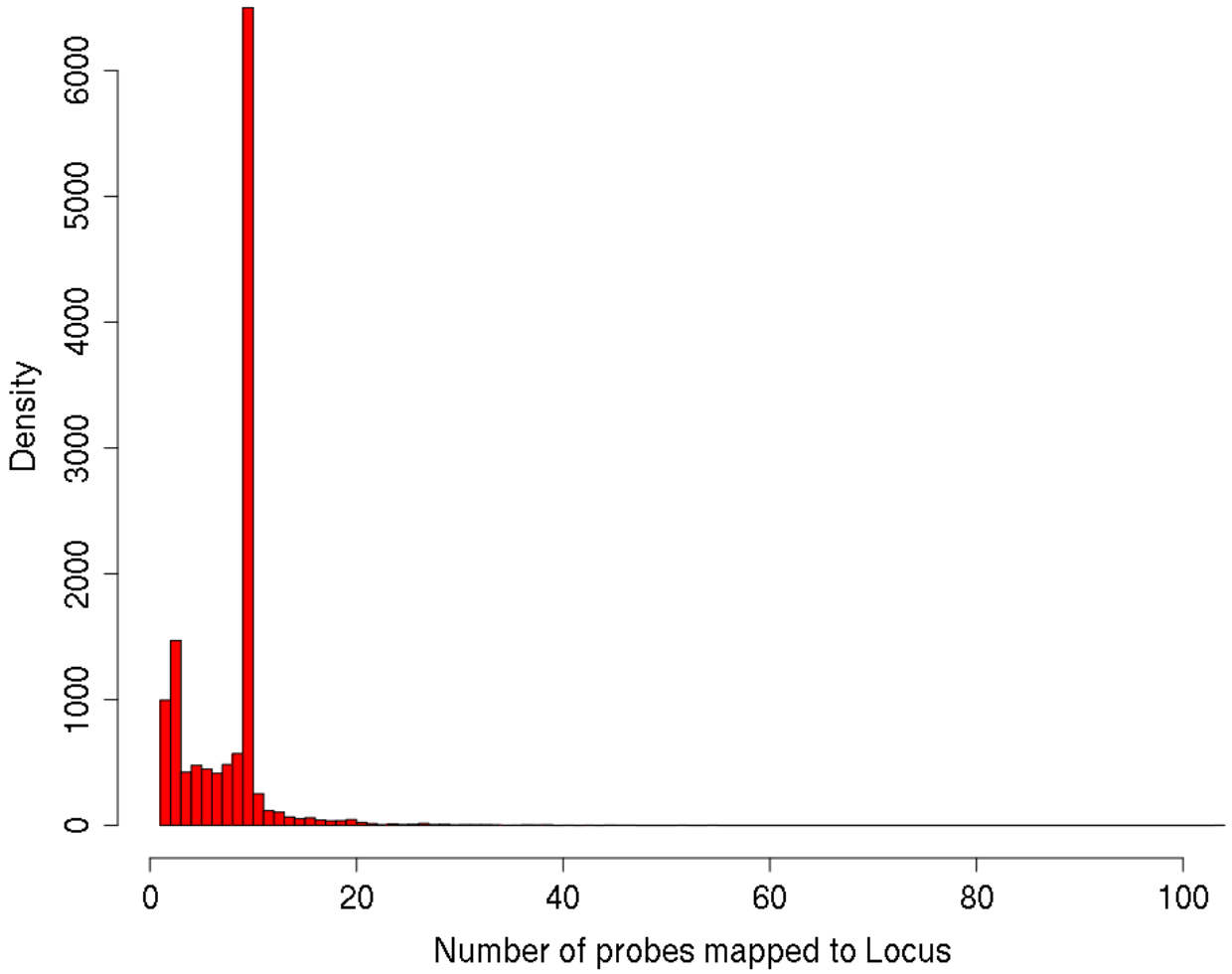
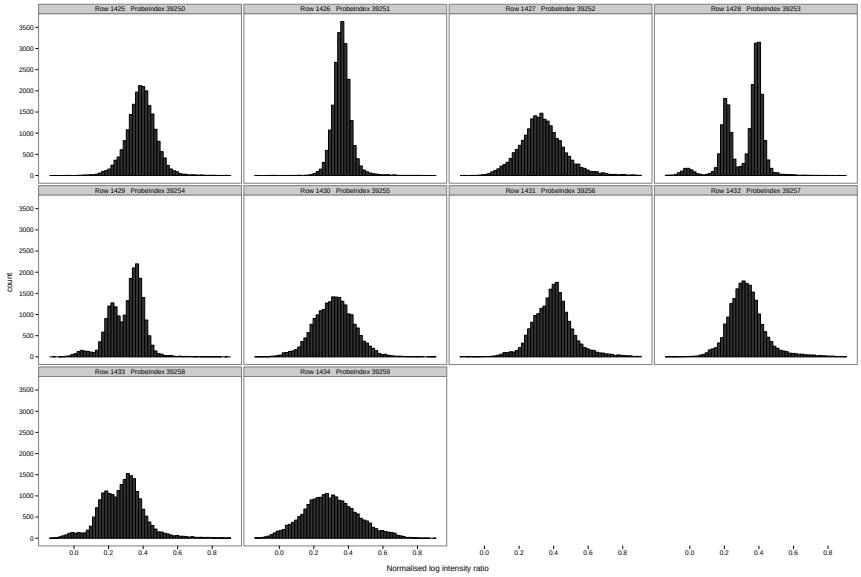
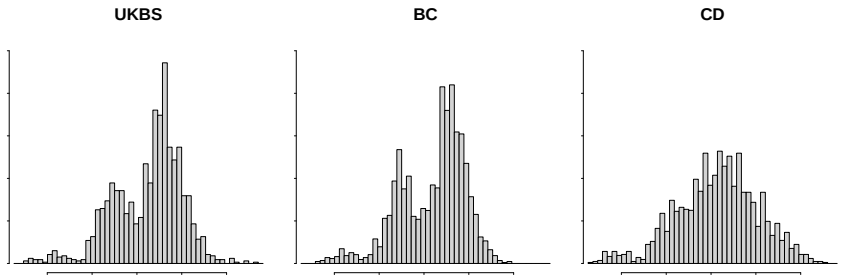


Figure 9: Distribution of the number of probes within each locus on the WTCCC CNV genotyping array. Two clear peaks can be seen at 3 and 10, representing the number of probes originally designed for exonic loci and other loci respectively. For the majority of CNVs that do not have the correct number of probes mapped, the number of probes is lower than expected due to filtering of probes implicit in the probe-to-CNV mapping algorithm. (This figure is referenced from the main text).

(a)



(b)



(c)

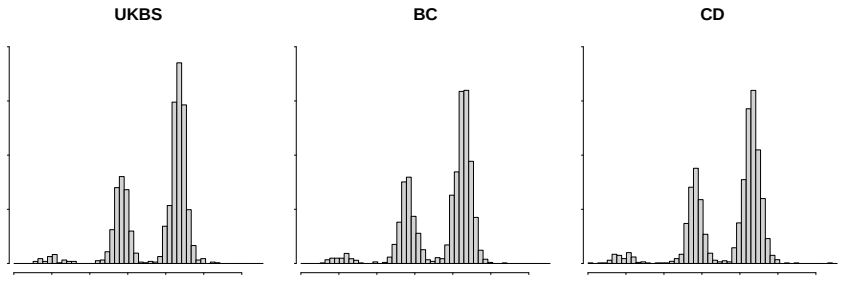


Figure 10: (a) Histograms of intensities across all individuals in the study separately for each probe for a particular CNV (CNVR3337.4). (b) Histograms for three collections of the normalised intensity for CNVR3337.4 summarised by the first principal component. (c) Histograms for same three collections of the normalised intensity for CNVR3337.4 summarised by the first principal component after use of probe variance scaling (PVS).

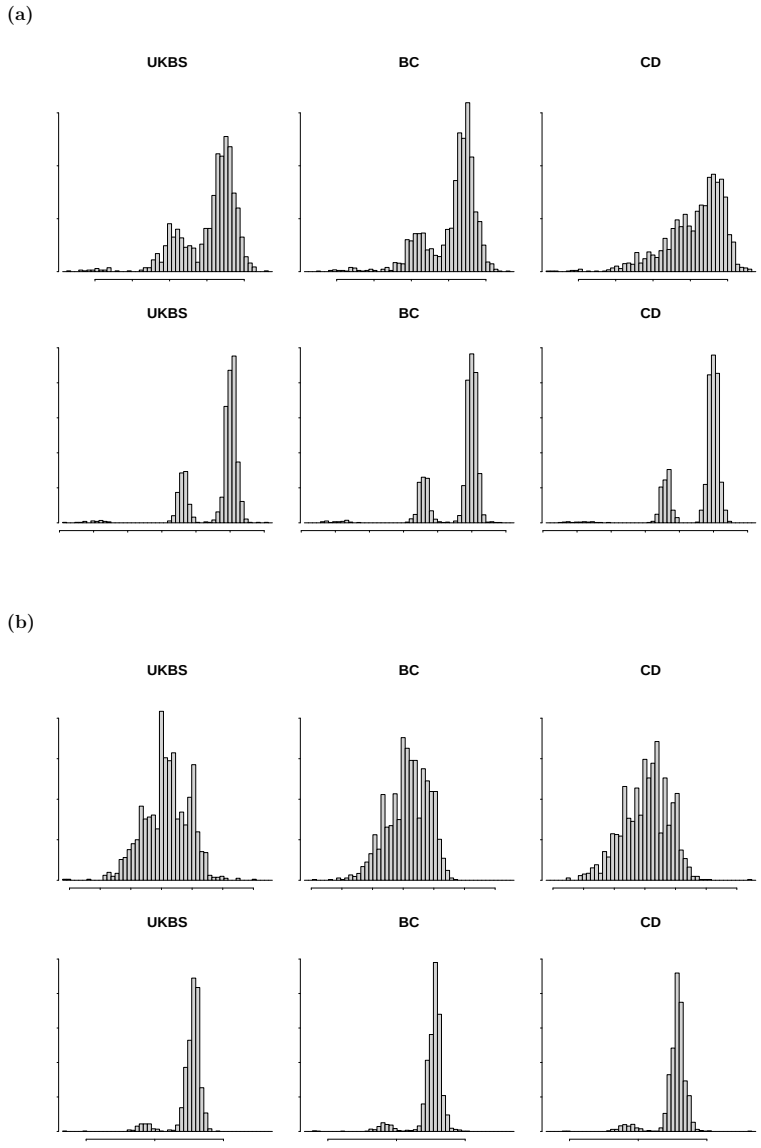


Figure 11: (a) Example of effects of different normalisations. The top row shows histograms of normalised intensity ratios for three collections for a CNV (CNVR2668.1) where the normalisation is that used in our standard pipeline ($log_2(QNorm(R)/QNorm(G) + 0.5)$). The bottom row shows this same CNV, but instead using just the raw log ratio ($log_2(R/G)$) with no normalisation. For this particular CNV, the calling algorithm can not reliably separate the classes when using the normalised data, whereas for the unnormalised data, the classes are clearly separated. (b) Example of effects of different probe summaries. The top row shows histograms of normalised intensity ratios for three collections for a CNV (CNVR4147.1) where the probes are summarised using the first principal component of the normalised intensity ratios. The bottom row shows this same CNV, but instead with the probes summarised by taking the mean value of the normalised intensity ratios of the probes. For this particular CNV, the calling algorithm cannot reliably separate the classes when using the PCA-summarised data, whereas for the mean-summarised data there is a clearly separated, rare second class.

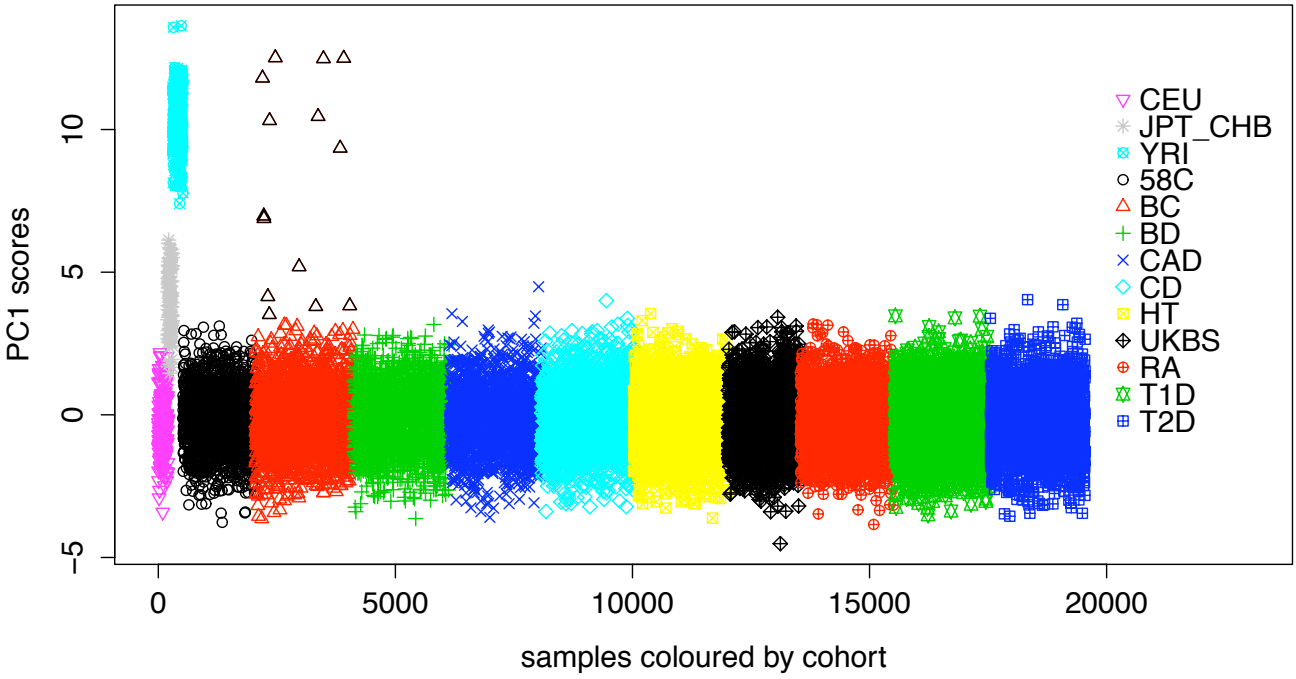


Figure 12: Results of an ancestry analysis applied to an initial set of CNV calls. The plot shows the scores from the 1st principal component (y-axis: PC1 Scores) for each WTCCC sample from a principal component analysis of the CNV calls. HapMap CEU, JPT, CHB and YRI samples are also shown. There is a point for each sample on the x-axis and samples are arranged by cohort with a legend detailing plotting symbols used for each cohort. The Black triangles indicate the 14 outlying BC samples.

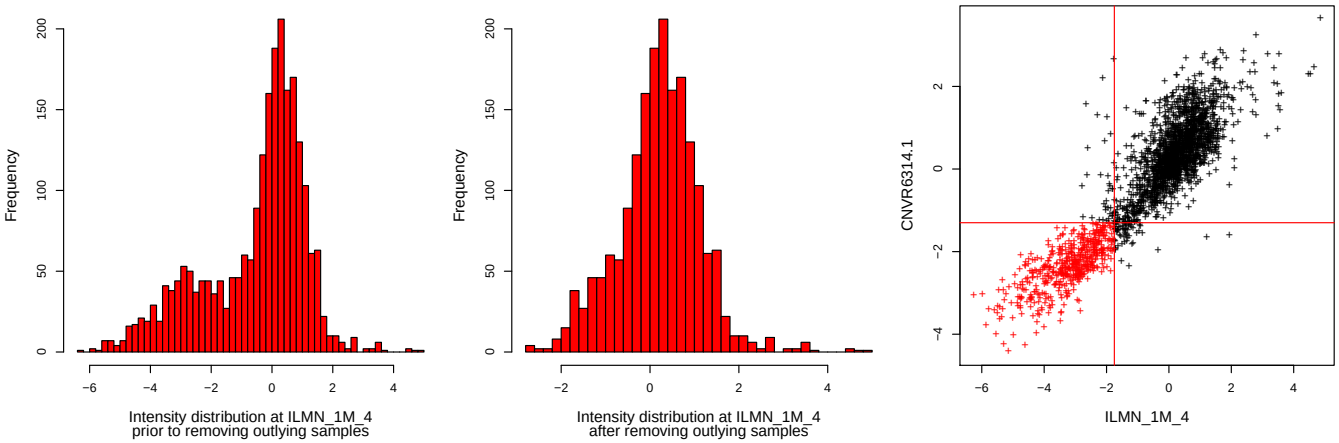


Figure 13: Illustration of the effect of the CAD outliers and the filtering used to remove them. These plots are illustrative for one particular CNV (ID: ILMN_1M_4). The first and second plots show the intensity distribution of CNV ILMN_1M_4 after (left plot) and before (middle plot) the exclusions respectively. The third plot (right) shows the bivariate plot of the two CNV intensity distributions used to determine the outlying samples (which deviate from the diagonal).

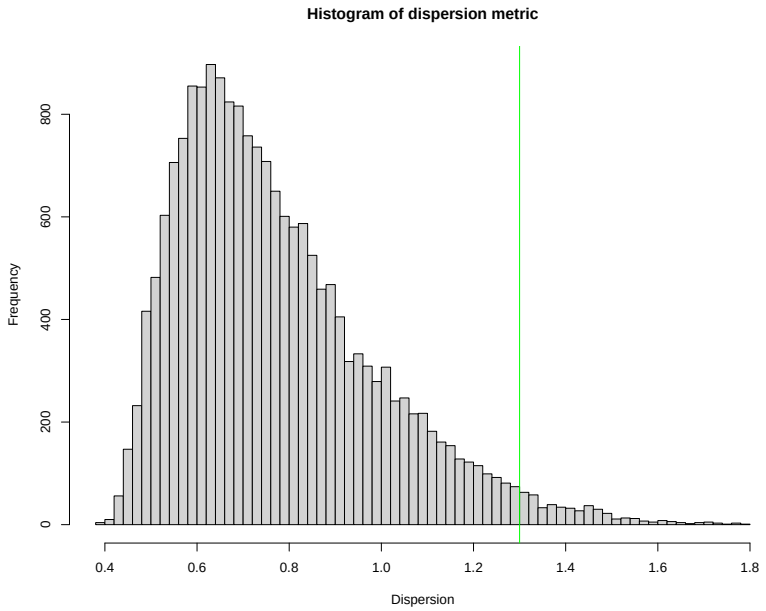


Figure 14: Histogram of dispersion metric for all post-calling samples. All samples included in the post-calling phase of the analysis are used in this frequency histogram. The green line at 1.3 is the cut-off that we used to identify poorly performing samples.

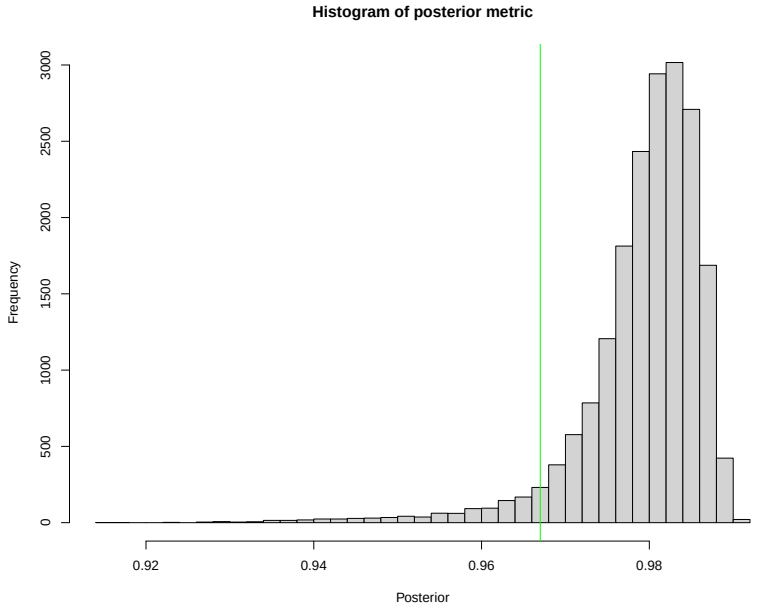


Figure 15: Histogram of posterior calling metric for all post-calling samples. All samples included in the post-calling phase of the analysis are used in this frequency histogram. The green line at 0.967 is the cut-off that we used to identify poorly performing samples.

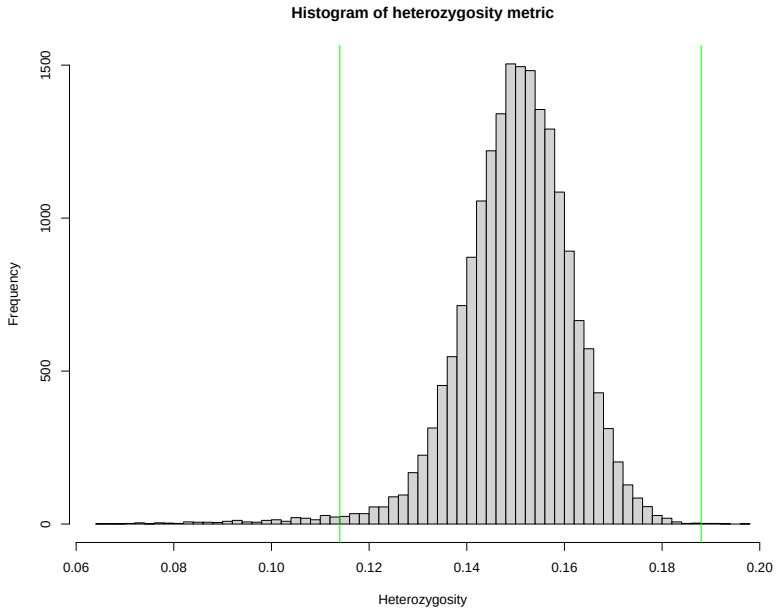


Figure 16: Histogram of heterozygosity metric for all post-calling samples. All samples included in the post-calling phase of the analysis are used in this frequency histogram. The green lines at 0.114 and 0.188 and are the cut-offs that we used to identify poorly performing samples.

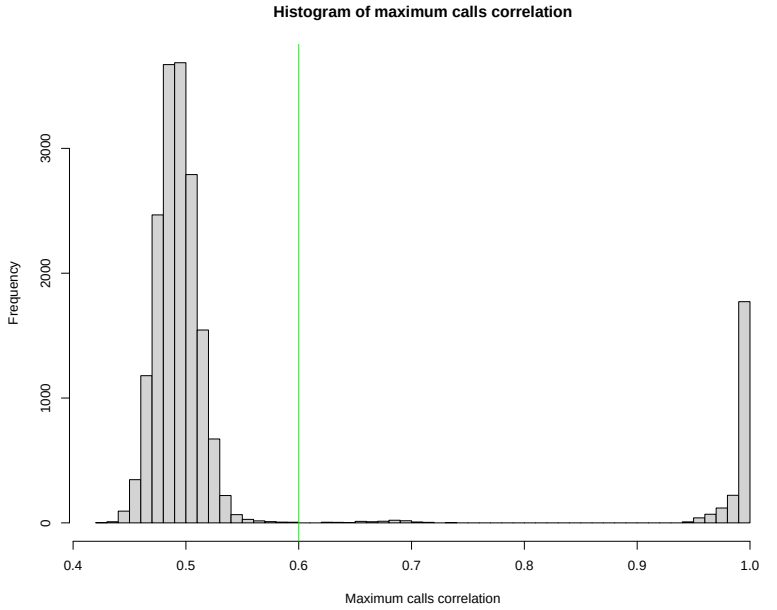


Figure 17: Histogram of maximum calls correlation with another sample for all post-calling samples. All samples included in the post-calling phase of the analysis are used in this frequency histogram. The green line at 0.6 is the cutoff used to identify whether two samples are closely related or the same individual. The points between 0.6 and 0.9 are assumed to be closely related samples, while those above 0.9 are assumed to be samples from the same individual.

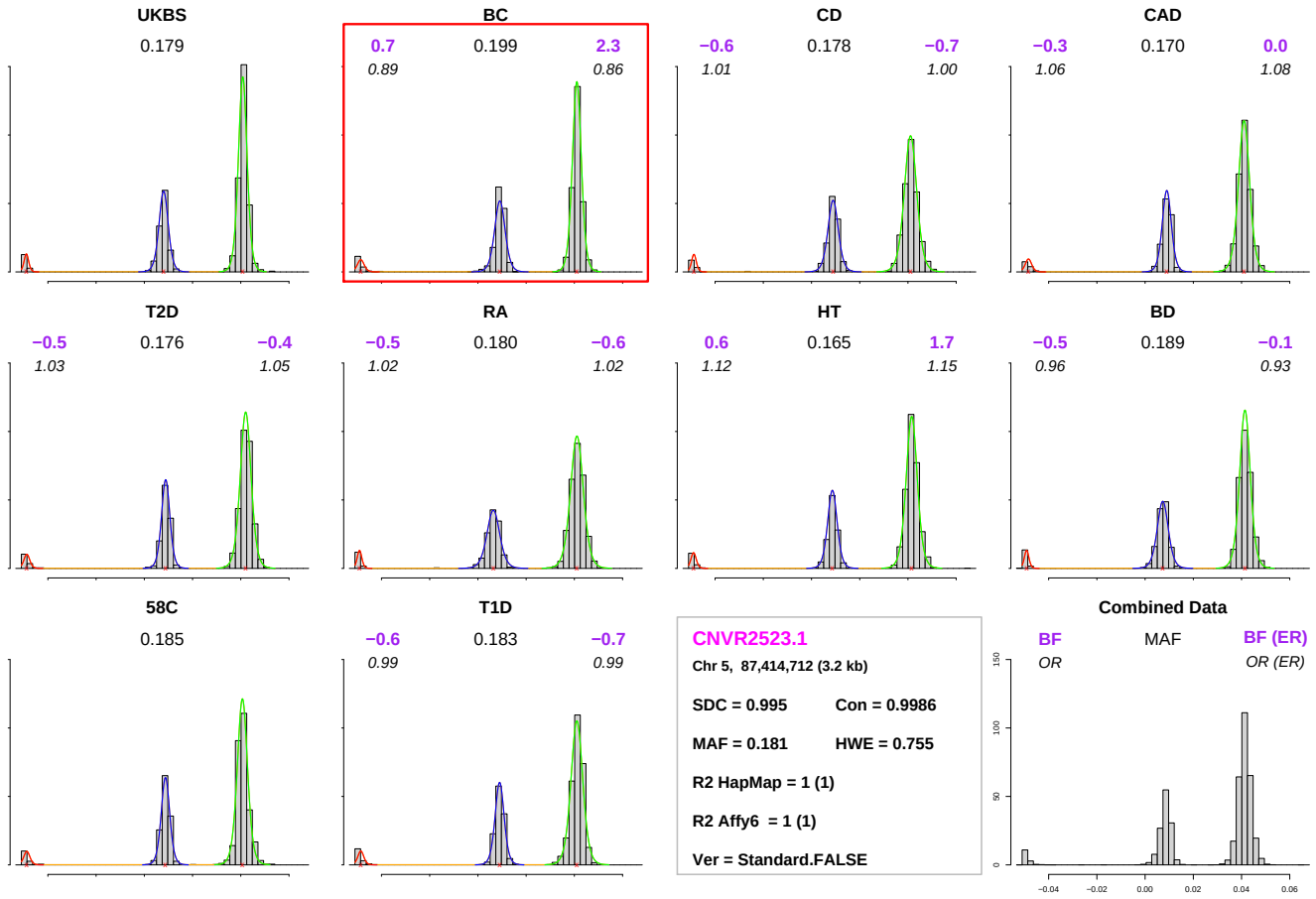


Figure 18: Plot of CNV calls and association testing results produced by the CNVCALL/CNVTEST approach. The CNV intensity distribution for each cohort are given in a separate sub-plot. The model fit (continuous coloured line) for each cohort is overlaid onto the intensity distributions. The means and variances of the CNV classes differ between cohorts but class proportions are set to the estimate of the overall class proportions. This aids visualization of effects since signals of association will show up as a difference between the model fit and the intensity histogram. The numbers at the top of each plot are as follows. Top left : log10 Bayes Factor (BF) for cohort versus the two control cohorts with associated estimate of the additive odds ratio for increasing copy number given below. Top right : log10 Bayes Factor for cohort versus the set of expanded reference panels (BF (ER)) (see Supplementary Table 10) with the associated estimate of the additive odds ratio for increasing copy number given below. Top middle : minor allele frequency estimate for that cohort. Case cohorts with either a BF or BF(ER) greater than 2.1 are highlighted by a red box around the subplot for that cohort. The CNV intensity distribution across all cohorts is given in the bottom right plot together with a legend for the 5 numbers included in each of the per-cohort subplots. The third subplot from the right on the bottom row gives information about the CNV shown in the rest of the subplots. The CNV names, chromosome, position and length of the CNV are given. The Strict Duplicate Concordance (SDC) is reported which is the ratio of all duplicate calls that agree divided by the total number of duplicates. The mean maximum posterior probability (or confidence (Con)) of the CNV genotype calls is given. The minor allele frequency (MAF) across all cohorts and the p-value for the test of Hardy-Weinberg Equilibrium (HWE) in all cohorts are shown. The maximum R^2 of the CNV to a HapMap SNP in a 1Mb region flanking the CNV is reported together with the maximum R^2 of the CNV to any HapMap SNP in brackets. The maximum R^2 of the CNV to a SNP on the Affymetrix 6.0 chip in a 1Mb region flanking the CNV is reported together with the maximum R^2 of the CNV to any Affymetrix 6.0 chip SNP in brackets. The version (Ver) of normalization used for the CNV is shown at the bottom of the subplot (see Supplementary Table 9 for more details). (This figure is referenced from the main text).

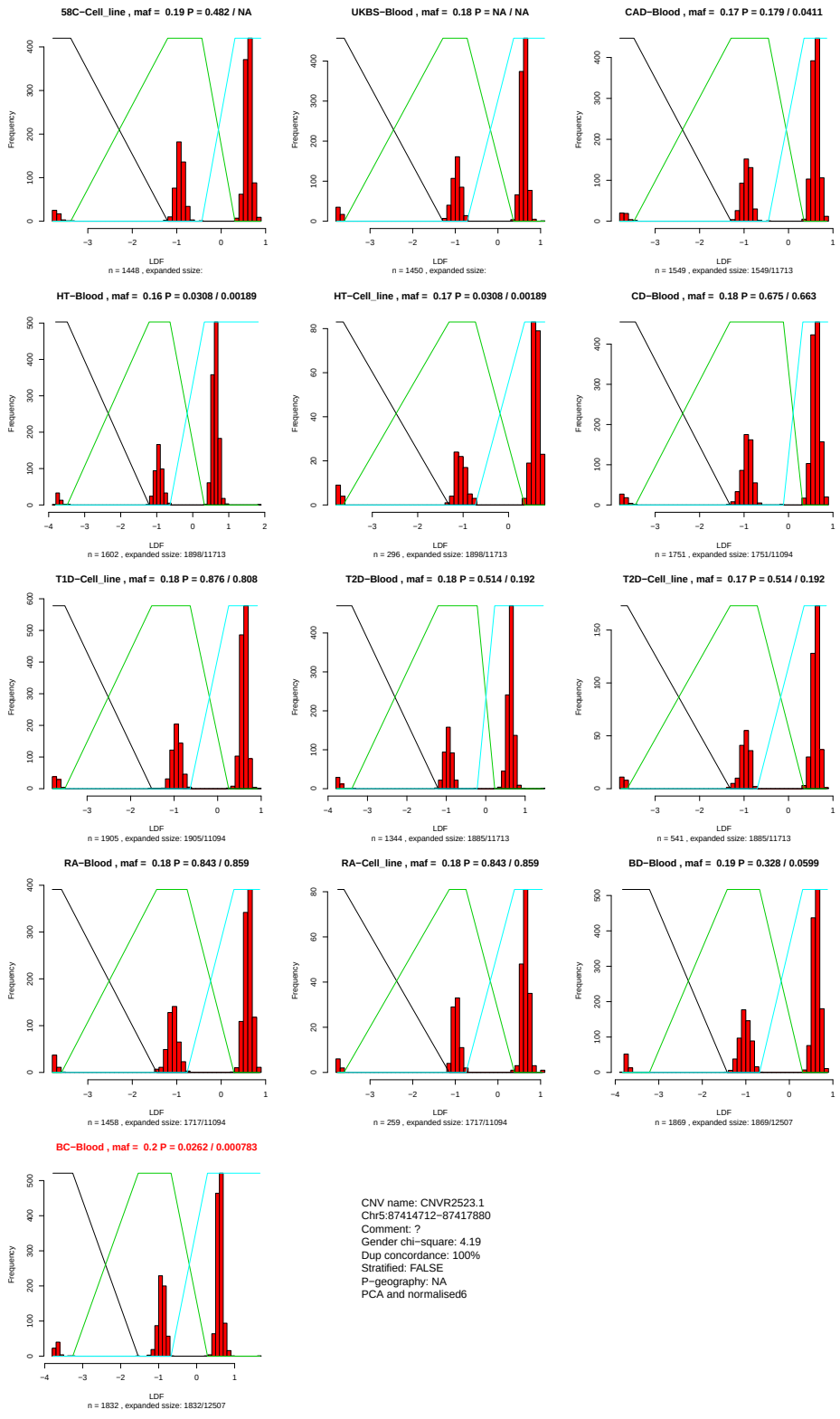


Figure 19: Plot of CNV calls and association testing results produced by the CNVtools approach. The CNV intensity distributions for each cohort are given in separate sub-plots. Where a cohort consists of samples from two sources (Blood derived DNA and Cell Line DNA) then two subplots are given, one each for samples from each source. The posterior probability for each CNV class is overlayed onto the intensity distributions.

The minor allele frequency estimate for the cohort and the p-values for the additive test of the cohort versus the two control cohorts and the expanded reference panels (see Supplementary Table 10) are given in the title of each subplot. If either or both of the 2 p-values (restricted or expanded) is < 0.001 then the title is colored red. The sample size of the cohort and the case and control sample sizes in the expanded reference analysis are given below each subplot. The bottom middle sub-plot first shows the CNV name and position. The "gender chi-square" category shows the goodness-of-fit test statistic for association between gender and copy number calls, which is distributed under the null of no association between gender and genotype as a chi-squared random variable with (n-1) degrees of freedom where n is the number of copy number classes. The "Dup Concordance" category shows the concordance at this CNV between the duplicate samples included in this study. This "Stratified: FALSE" indicates that regions of origin were not included as covariates. The P-value testing association between region and copy number call, denoted as "P-geography", is only shown when a stratified test of association was used. The last row shows the normalization and the summary method (either an initial PCA or mean signal summary across CNV probes (see Supplementary Table 9) , always followed by a linear discriminant analysis step). (This figure is referenced from the main text).

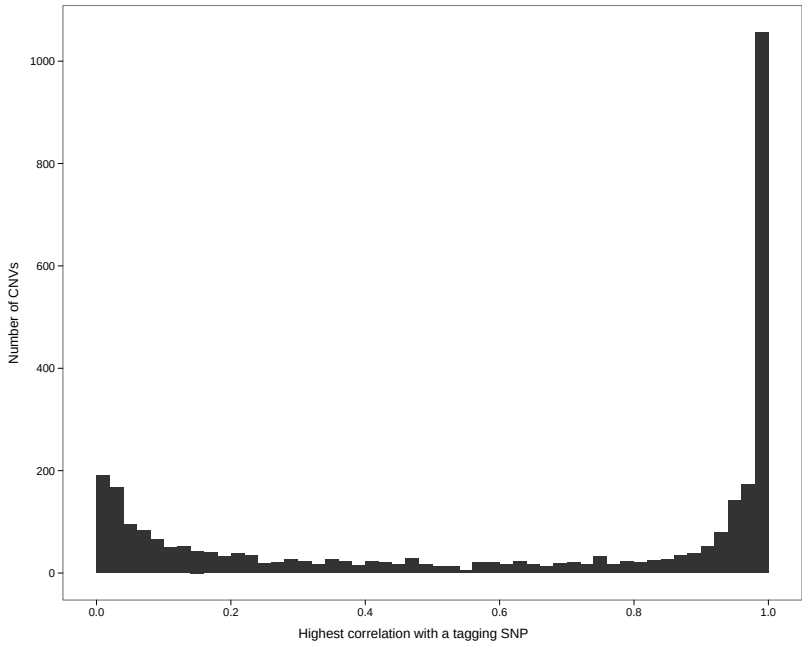


Figure 20: Histogram of maximum correlation r^2 between each CNV and a SNP within 1MB of the ends of that CNV. The histogram represents all 3,188 autosomal CNVs (i.e. excluding X-chromosome CNVs and CNVs on novel insertions) that passed QC metrics using the CNVCALL approach. (This figure is referenced from the main text).

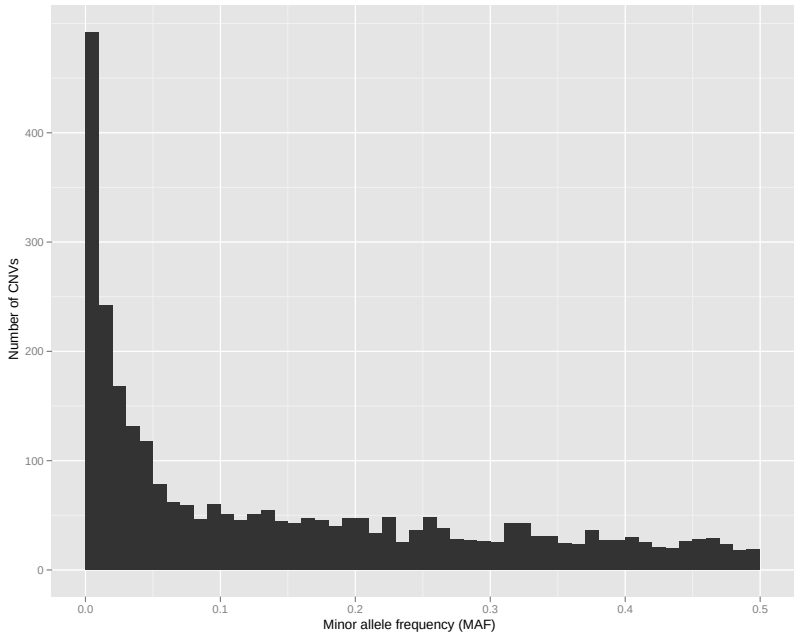


Figure 21: Histogram of minor allele frequency for bi-allelic CNVs. Bi-allelic CNVs were identified as those CNVs that passed QC, had 2 or 3 classes, and had a Hardy-Weinberg equilibrium p-value of greater than 1×10^{-7}. (This figure is referenced from the main text).

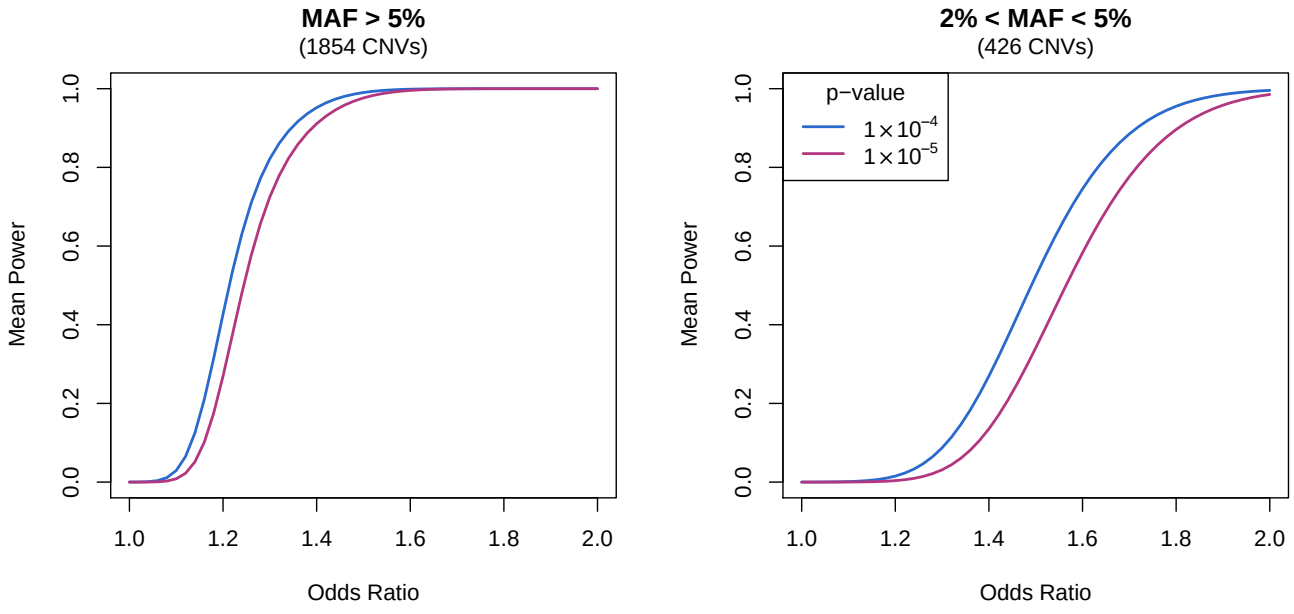


Figure 22: Plot showing the mean power for common (left) and rare (right) CNVs. Only non-duplicate, well-separated autosomal CNVs which were called with either 2 or 3 classes were used. The x-axis of the plots give the odds ratio and the y-axis shows the mean power. Power at two p-value thresholds are given : 1×10^{-4} (blue) and 1×10^{-5} (purple). (This figure is referenced from the main text).

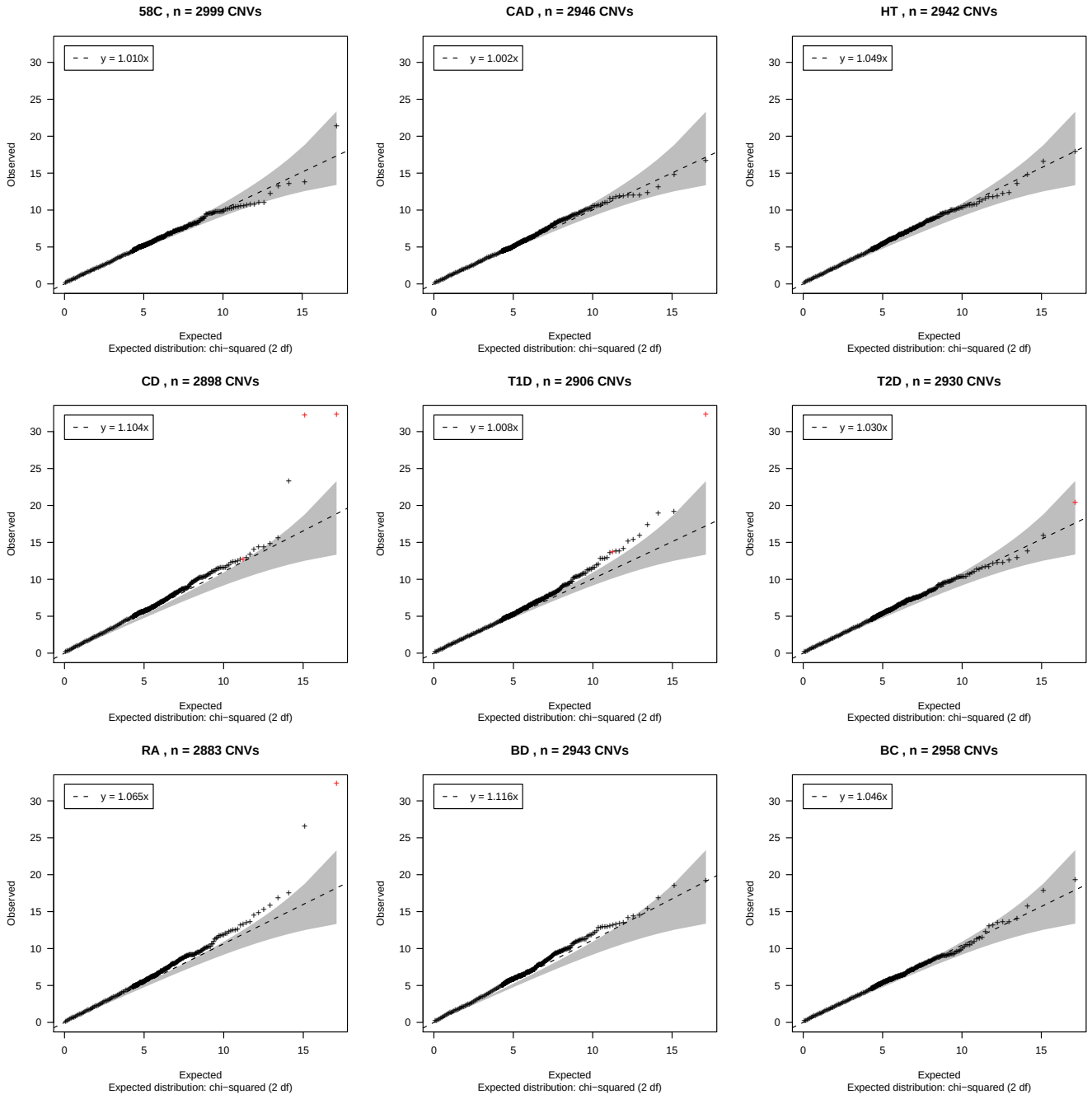


Figure 23: Quantile-quantile plots comparing the expected versus observed distribution of -2log(p), where p is the p-value for the one degree-of-freedom linear trend test of association. Under the null hypothesis of no association p is uniformly distributed between 0 and 1, and therefore -2log(p) is distributed as chi-square on two degrees of freedom. CNVs included in these plots were filtered on the basis of a clustering quality score (see Section 6.1 for details, numbers of CNVs for each disease shown on the plot) and manual inspection of the most significant associations. Gender related artefacts and CNVs in the HLA for autoimmune diseases were removed. The dashed line in each plot is based on an estimate of possible over- or under-dispersion of the test statistics, and the difference between the slope of the line (often denoted by λ) and unity gives a numerical estimate of overdispersion. (This figure is referenced from the main text).

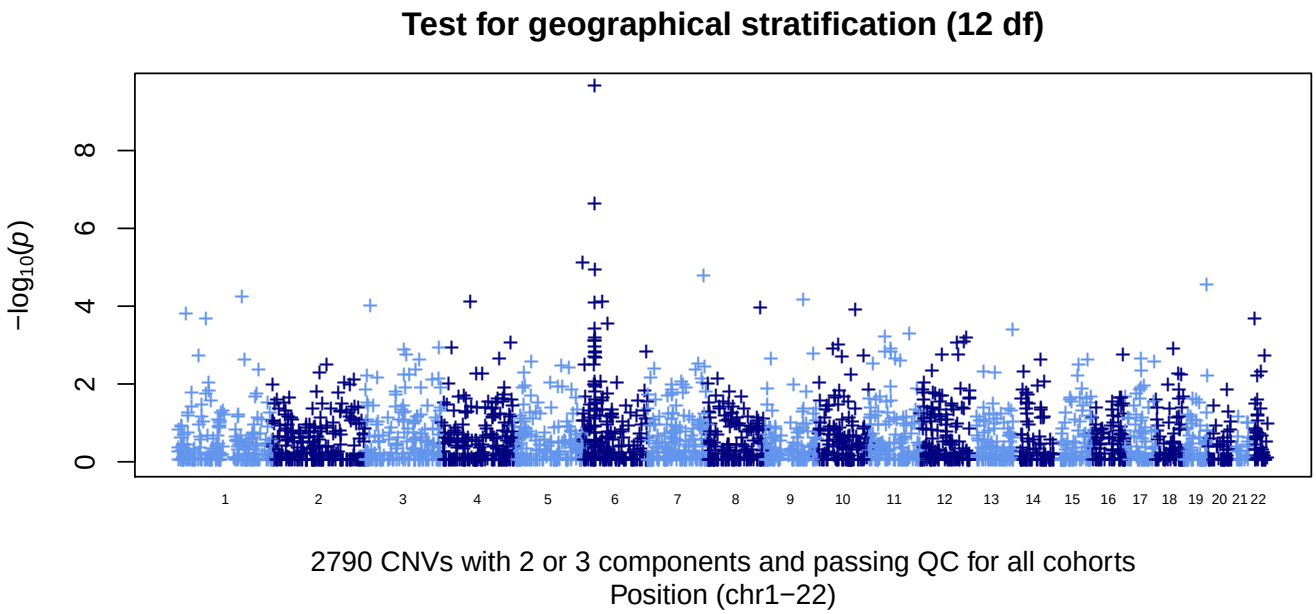


Figure 24: Manhattan plot showing the results of the test for geographic stratification. The y-axis shows the -log10(p-value) for the 12 degrees-of-freedom test of association between CNV genotype and the region of origin of each sample in the study. The x-axis shows the p-values for each CNV arranged by autosome. (This figure is referenced from the on-line methods).

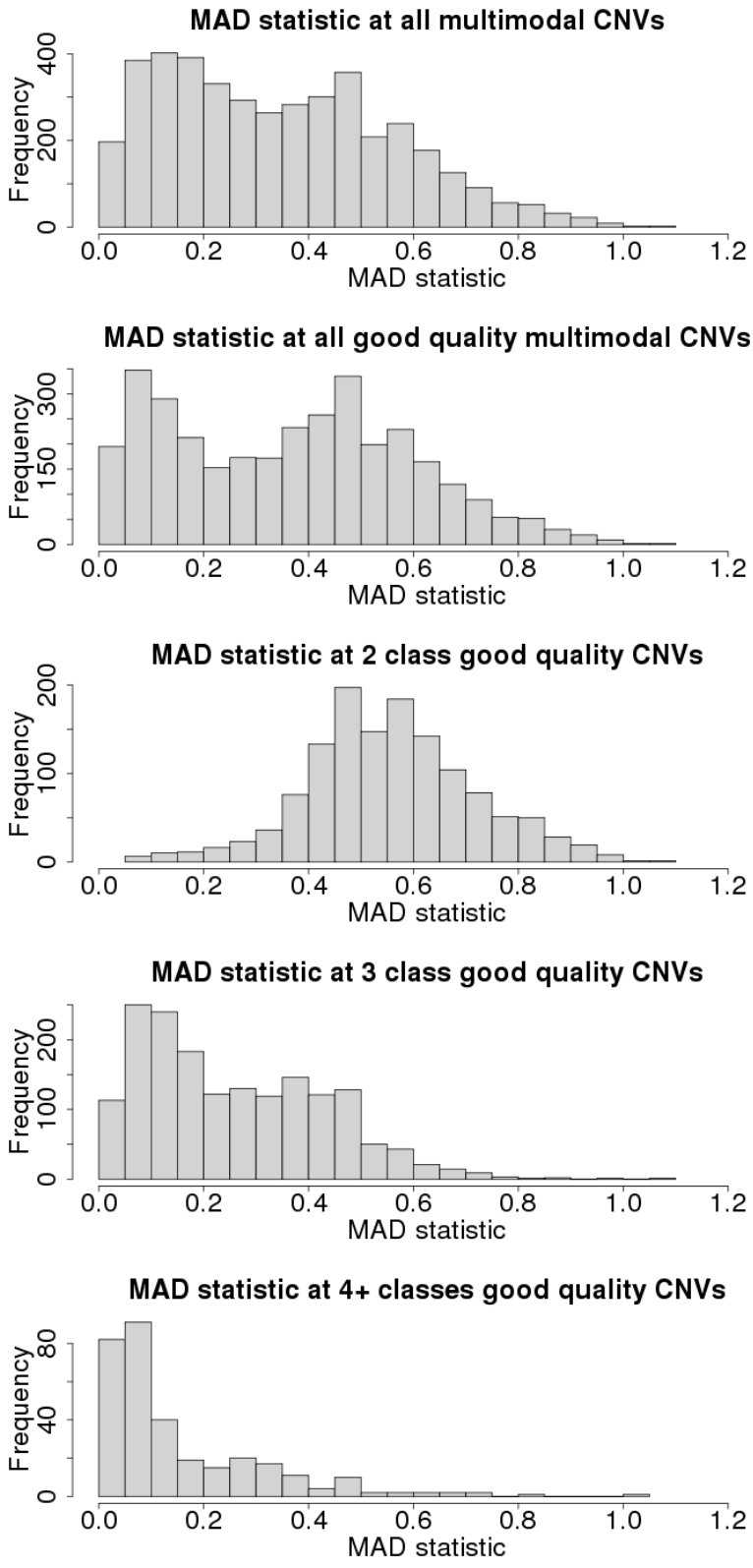


Figure 25: Distribution of the MAD (median absolute deviation) statistic at multi class CNVs. The top plot shows the distribution across all CNVs called with more than one class. The second row shows the distribution when only CNVs with good quality calls are used. This is then split up into 3 categories according to the number of classes called. The MAD statistic decreases as the number of classes increases which shows that it is well correlated with the amount of polymorphism in the data. The number of classes is calculated from the results of our best pipeline (described elsewhere) and the MAD statistic is calculated based on that data.

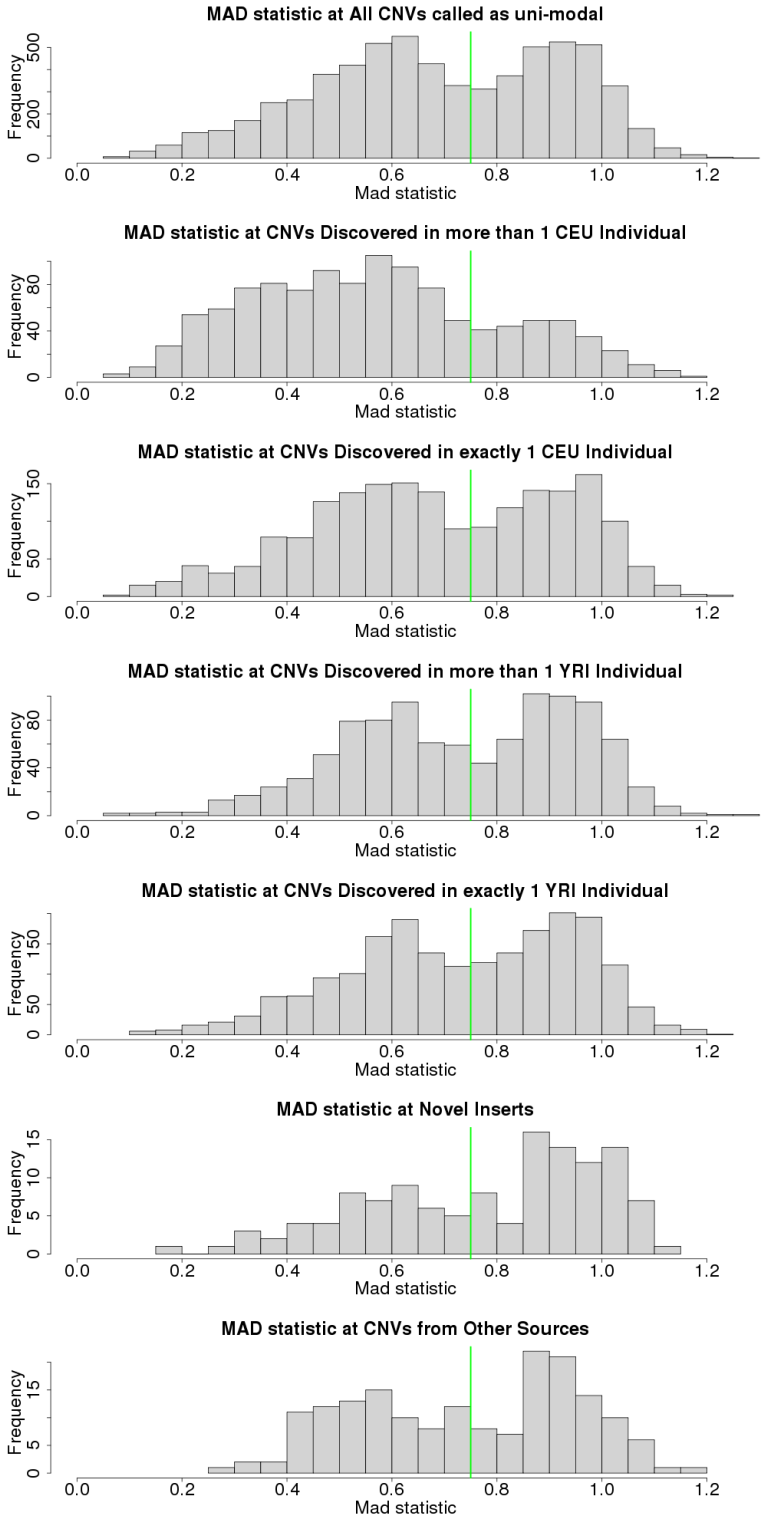


Figure 26: Distribution of the MAD (median absolute deviation) statistic at single class CNVs. The top plot shows the distribution across all single-class CNVs. Then this is stratified by properties of the discovery experiment. The second row shows the distribution for those CNVs discovered in multiple CEU samples, then those discovered in exactly one CEU sample, of those CNVs not discovered in any CEU samples the remaining CNVs are split according to whether they were discovered in more than one YRI sample or exactly one YRI sample. The single-class CNVs correspond to those for which no pipeline gave good quality calls with more than class, however the statistic calculated for the monomorphic CNVs is based on the signal in the standard pipeline.

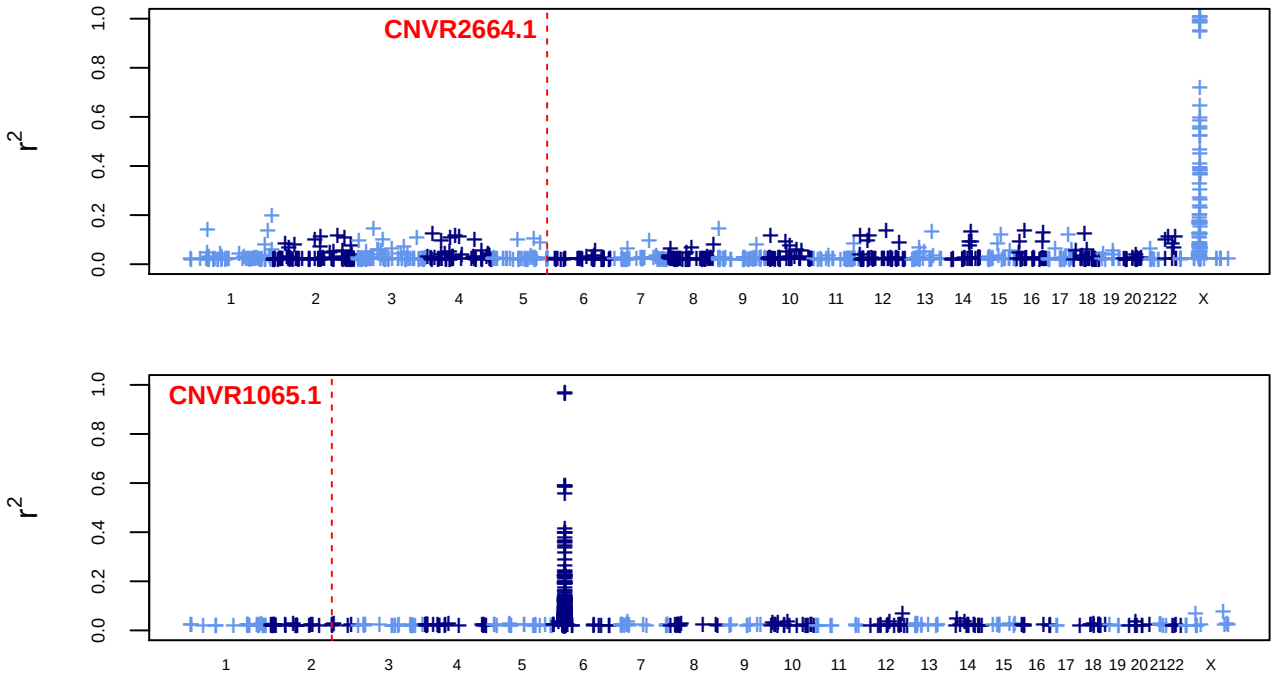


Figure 27: Correlations between two dispersed duplications, CNVR2664.1 and CNV1065.1, and SNPs across the genome. The red dashed vertical lines indicate the genomic location in the reference sequence to which the probes in the CNV uniquely map. In each case, the SNP-tagging results show the variation to be elsewhere (on the X-chromosome and in the HLA, respectively). (This figure is referenced from the main text).

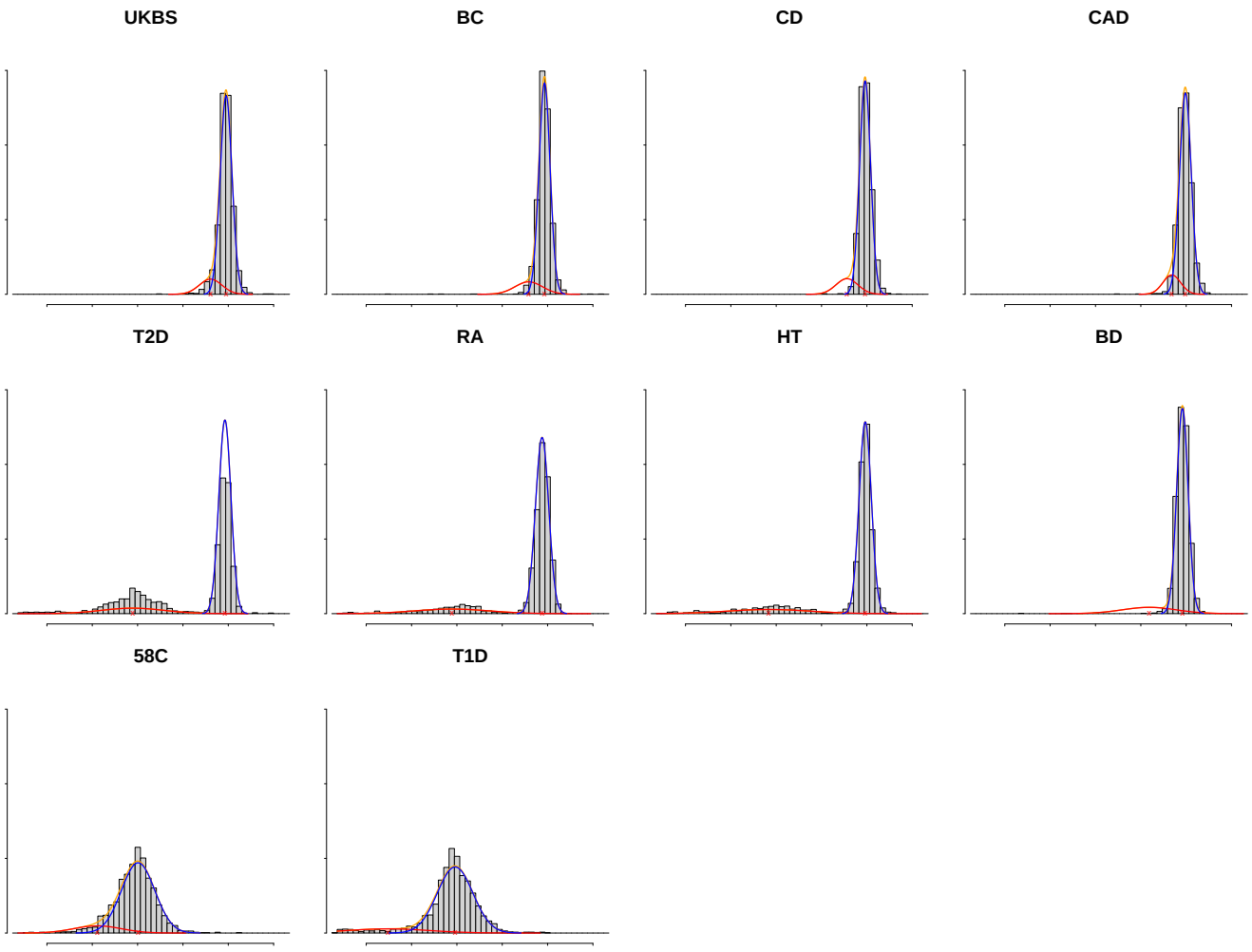


Figure 28: Example of a DNA source effect. The 10 plots show histograms of normalised intensity ratios for all 10 collections (8 case and 2 control) for CNVR866.8. It can be seen that the four collections in the top row appear to have one class, and the two collections in the bottom row also appear to have one class, but with quite different mean normalised intensity ratios. Furthermore, T2D, RA and HT collections appear to have two classes. Note that the DNA samples from the collections in the top row and BD were all derived from blood, the DNA samples in the collections in the bottom row were all derived from cell-lines, and the DNA samples from the T2D, RA and HT collections were derived from both blood and cell-lines.

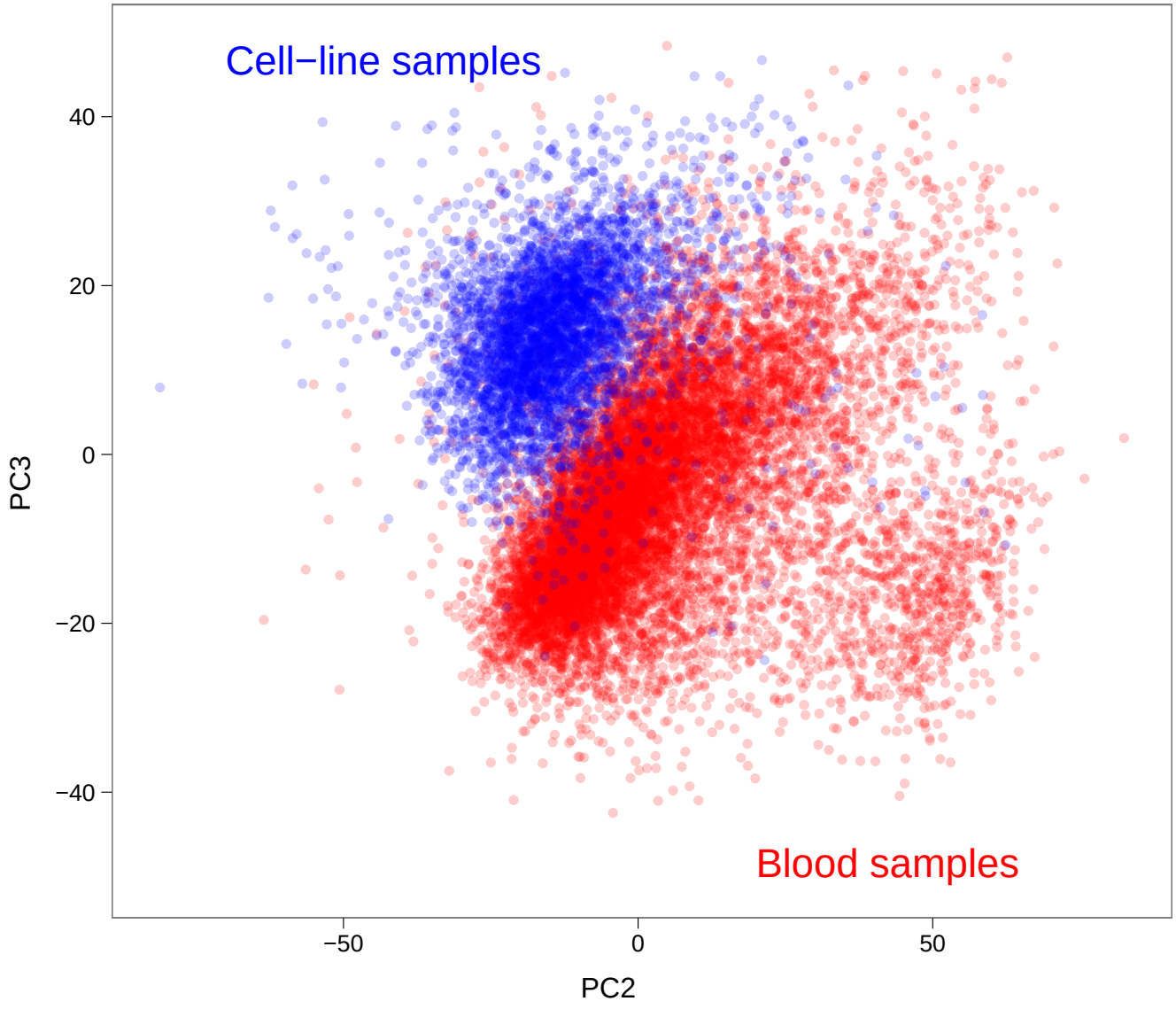


Figure 29: DNA source effect after removal of Ig genes. This plot was created using all samples post QC from all 10 collections using data from all CNVs with the exception of those within 1MB of known Immunoglobulin genes. Each point represents one sample, with the points coloured according to whether that sample was derived from blood (red) or cell-lines (blue). The two axes represent the second and third principal components of the CNV-level (summarised from the probes using the mean of the probes) normalised log-ratio intensity data. The two overlapping clusters amongst the blood-derived DNA samples do not appear to relate simply to particular variables such as collection or DNA extraction technique. (This figure is referenced from the main text).

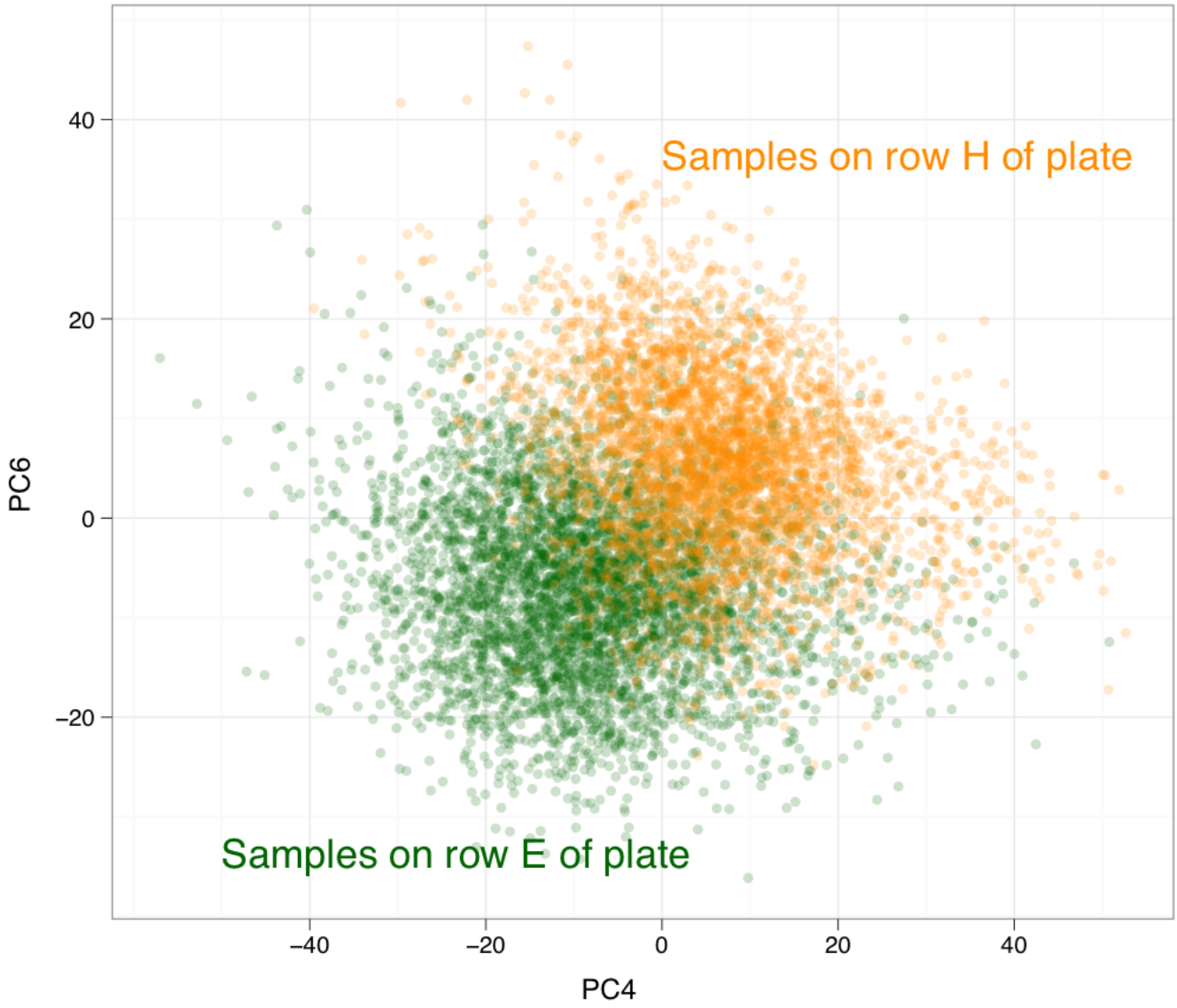


Figure 30: Cluster plot of plate row effect. This plot was created using samples from all 10 collections that were on the two extreme rows of the plate (row E and row H), and all CNVs. Each point represents one sample, with the points coloured according to whether that sample was on row E (green) or row H (orange). The two axes represent the fourth and sixth principal components of the CNV-level (summarised from the probes using the mean of the probes) normalised log-ratio intensity data. (This figure is referenced from the main text).

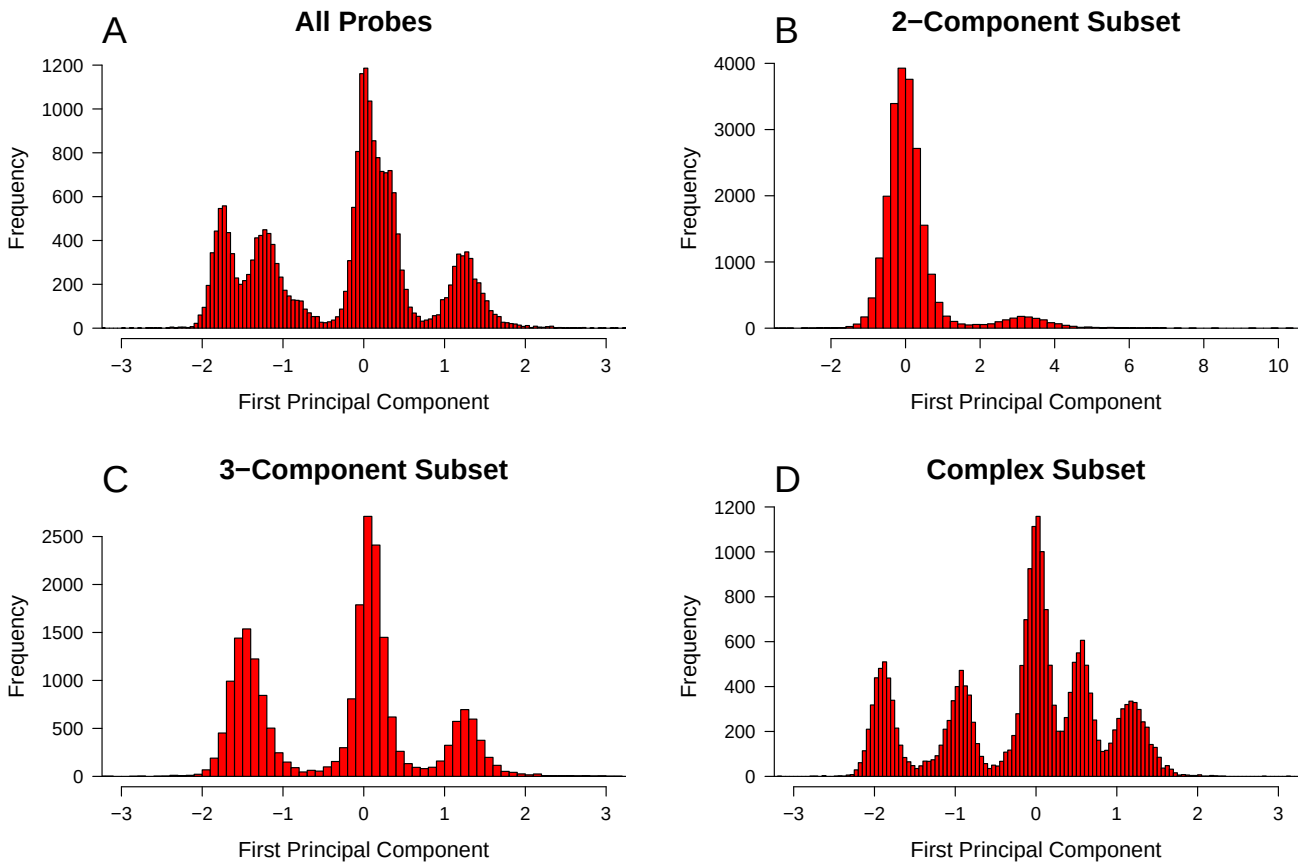


Figure 31: Histograms of the first principal component for CNVR5583.1 for multiple distinct probe subsets within the CNV region. A) Cluster separation is poor for CNVR5583.1 when using all 10 probes within the breakpoints defined from the GSV study. However, the signal at each of the 10 probes within the breakpoints specified for CNVR5583.1 either appeared to be non-polymorphic (3 probes at the boundaries of the region), or fall into one of three distinct cluster patterns: B) a 2-component CNV subset (albeit with a very rare third component) consisting of 2 contiguous probes; C) a clear 3-component CNV subset consisting of 4 contiguous probes; D) a much more complex CNV subset with 5 distinct copy-number classes detected by a single probe located between the two larger subsets.

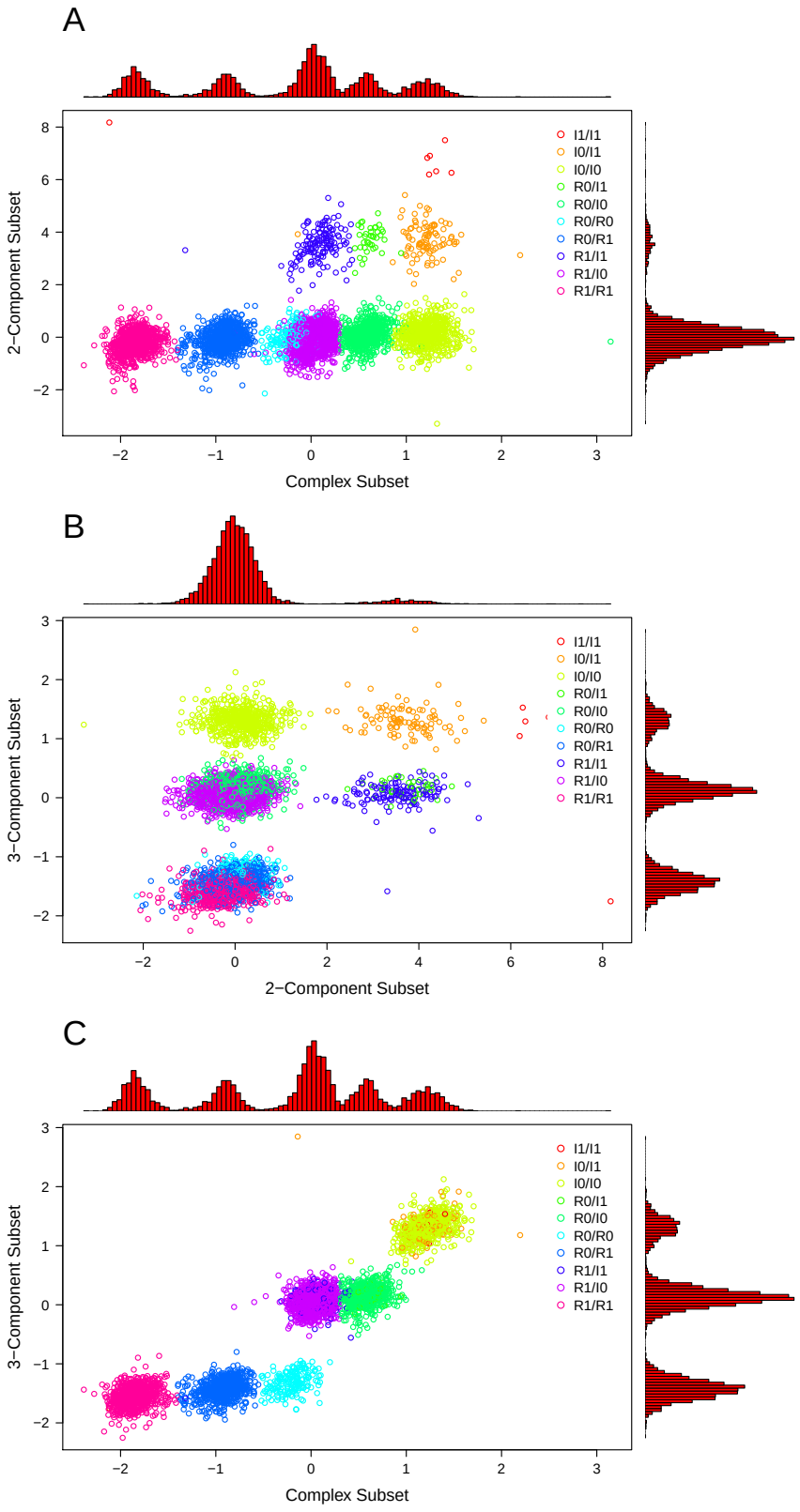


Figure 32: Suspected haplotype combinations account for copy number classes detected using CNV genotyping array for CNVR5583.1. The three subsets of probe signal detected within the CNVR5583.1 region can be accounted for by 4 possible haplotypes: R0, R1, I0 and I1. Sequencing of the region suggests that these haplotypes may occur as a result of a combination of deletion and inversion events, although further sequencing is required to fully characterize this complex region. These four haplotypes can occur occur in 10 different combinations, which are clearly distinguishable particularly in A.

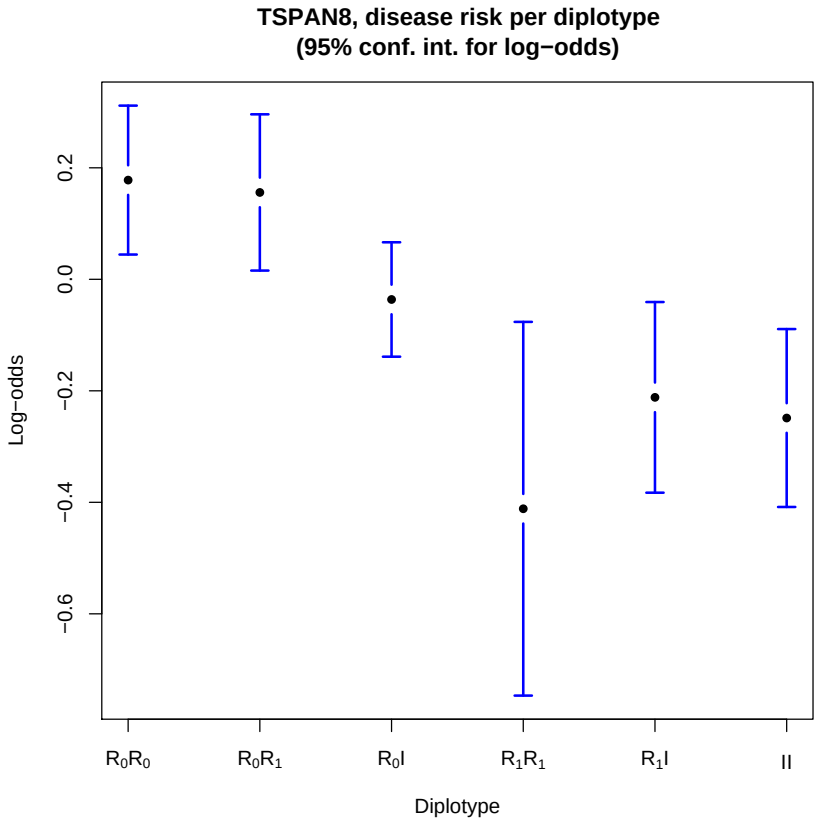


Figure 33: Estimated log-odds and 95% confidence intervals for the T2D association in each diplotype category. The R/I haplotypes are defined using the clearest three-component CNV (Supplementary Figure 31C). For individuals with either 1 or 2 copies of the R haplotype, subgroups R0 and R1 are defined based on the CNV in Figure 31D. Further subdivisions of the I haplotype between I0 and I1 (based on the CNV in Supplementary Figure 31B) are not shown here.

14 Tables

	Agilent	Illumina	NimbleGen
Deletion	0.68	0.53	0.51
Duplication	0.64	0.64	0.27
Multiallelic	0.44	0.20	0.08

Table 1: Proportion of successfully clustered polymorphisms by class of CNV for each platform used in the pilot study.

Collection	58C	UKBS	BC	BD	CAD	CD	HT	RA	T1D	T2D
Total samples	1500	1500	2000	2007	2000	2016	2000	2007	2015	2005
Samples studied in WTCCC1	1367	1416	0	1808	1871	1568	1931	1823	1918	1771
Samples not studied in WTCCC1	133	84	2000	199	129	448	69	184	97	234
Intended duplicates	47	94	94*	47	47	47	47	47	47	47

Table 2: Sample totals for each cohort. Number of samples for each cohort that were sent for CNV genotyping with breakdown according to sample overlap with those typed for SNPs in WTCCC1. The number of planned duplicates is also shown. *: 47 blood and same 47 as cell-line.

Class	Source	Available	Attempted	Included	% success	% included loci
Control loci	chrX	10	10	10	100.00%	0.09%
CNV loci	WTCCC1 CNVs	18	18	18	100.00%	0.15%
	GSV CNV map	10865	10329	9722	94.10%	83.62%
	Affymetrix 6.0	85	85	83	97.60%	0.71%
	Illumina_1M	85	85	82	96.50%	0.71%
	WTCCC1	228	228	228	100.00%	1.96%
	Novel sequences	292	292	292	100.00%	2.51%
Candidate genes	Exons	994	994	918	92.40%	7.90%
Validation loci	WTCCC1	297	297	274	92.30%	2.36%
Total		12874	12338	11627	94.20%	100%

Table 3: Summary of the number of loci targeted on the array design from each source of loci.

Gene	Collection
CAD	*ACE*, *ACE2*, *CMA1*, *F12*, *REN*, *SERPINA8*
T1D	*AIRE*, *CD226*, *CLEC16A*, *FOXP3*, *IL18RAP*, *IL2*, *IL2RA*, *PTPN2*, *SH2B3*, *TAGAP*
BD	*ANK3*, *CACNA1C*, *DISC1*, *NRXN1*, *ZNF804A*
BC	*ATM*, *BRCA1*, *BRCA2*, *BRIP1*, *CDH1*, *CHEK2*, *PALB2*, *PTEN*, *RAD50*, *STK11*, *TP53*
RA	*CCL3L1*,*CCL4L*, *CCL4L2*, *CD40*, *DEFB4*, *FCGR3B*, *KIF5A*, *TRAF1*
T2D	*CDKAL1*,*CDKN2A*, *CDKN2B*, *FTO*, *GLUD1*, *HHEX*, *IDE*, *JAZF1*, *KIF11*, *PPARG*, *TCF2*
UKBS	*GP1BA*, *GP1BB* ,*GP5*, *GP6*, *GP9*, *VWF*
CD	*IL12B*, *IL23R*, *NKX2-3*, *NOD2*, *PTPN2*, *SLC22A5*
HT	*PODN*, *SLC2A9*, *SLC6A2*, *SLC9A3*, *WNK1*

Table 4: List of selected candidate genes for each cohort.

Mapped Probes	All Loci	CNVs	Exons	Novel Inserts	Control Regions
1	559	504	54	1	0
2	437	389	48	0	0
3	1469	390	1078	0	1
4	423	422	1	0	0
5	477	476	1	0	0
6	448	446	1	1	0
7	414	414	0	0	0
8	483	483	0	0	0
9	571	571	0	0	0
10	6500	6208	2	290	0
11-15	592	588	0	0	4
16-20	221	215	6	0	0
21-25	61	59	0	0	2
26-30	43	42	0	0	1
31+	42	39	1	0	2
Total	12740	11246	1192	292	10

Table 5: **Number of probes per locus** Probes on the array were mapped to each locus of interest based on their genomic coordinates and the number of additional loci to which the probe mapped. The subset of probes mapping to the least number of additional loci was chosen to ensure that summary data over probes was, as far as possible, representive of a single event only as much as possible. The majority of CNVs were targeted by the probes specifically designed for that purpose (3 probes for exon regions, 10 probes for other loci). However, there were a large number of loci for which the summary data were taken over a smaller subset of probes due to the presence of probes mapping to multiple locations within the genome. There were also a smaller number of loci for which the probe subset taken was larger than the designed set of probes, including 39 CNVs queried by more than 30 probes in total.

C2078	95040533	Male
C2141	950515172	Male
C2153	950515226	Male
C2173	95061994	Male
C2175	95061996	Male
C2188	960425324	Male
C2159	950515238	Male
C2184	960425320	Female
C2142	950515190	Male
C2151	950515201	Male

Table 6: Identification numbers of the 10 DNA samples from the European Collection of Animal Cell Cultures (ECACC) that were pooled to form the reference DNA sample.

		Number of classes using CNVCALL				
		2	3	4	5+	Total
Number	1	117	16	3	2	138
of	2	1,027	150	15	0	1,192
classes	3	162	1,504	203	14	1,883
using	4	5	22	39	18	84
CNVtools	5+	10	5	4	23	42
	Total	1,321	1,697	264	57	3,339

Table 7: Estimated numbers of classes for both calling algorithms (CNVtools and CNVCALL) for the 3,339 autosomal CNVs passing QC (which necessarily entails more than one class for CNVCALL). Shaded cells correspond to CNVs where numbers of classes agree between the two approaches. (This table is referenced from the on-line methods).

		Samples excluded before calling									Excluded before testing			
Collection	Total samples sent for assay	Supplier error	Sample handling error	Duplicate in multiple cohorts	Non European ancestry	Mixed sample	Low signal	DLRS fail	Initial calling quality metric fail	Total pre-calling exclusions	Post-calling quality metric fail	Duplicates and close relatives	Total samples used in CNV association testing	Proportion of females in sample tested for CNV association
UKBS	1659	8	0	0	0	47	3	15	28	101	71	37	1450	52%
58C	1671	2	0	0	0	0	3	36	22	63	79	81	1448	48%
BC	2134	3	0	1	14	0	12	39	36	105	123	74	1832	100%
BD	2134	27	0	2	0	0	4	20	50	103	95	67	1869	62%
CAD	2345	13	2	4	0	47	6	190	9	676*	67	53	1549	22%
CD	2322	27	1	0	11	47	29	158	63	336	121	114	1751	60%
HT	2190	4	0	5	0	0	5	69	18	101	116	75	1898	60%
RA	2254	46	3	1	1	46	5	41	120	263	202	72	1717	74%
T1D	2205	2	2	1	0	0	1	73	15	94	134	72	1905	49%
T2D	2186	17	7	4	0	2	4	39	48	121	91	89	1885	42%
Total	21100	149	15	18	26	189	72	680	409	1963	1099	734	17304	58%

Table 8: Numbers and genders of samples used in the CNV association analysis and the breakdown of reasons for exclusion from analysis of samples sent for laboratory assay but not used in analysis. Note that this includes all samples analysed during this experiment including those repeated. Supplier error refers to a sample-related error originating in the relevant disease group and that was identified after sending for assay but before analysis of the data (for example, accidental duplicates or evidence that the sample was not the same as that specified in the sample manifest). Sample handling error refers to an error in the central processing of the DNA samples at the Sanger laboratory. Duplicate in multiple cohorts refers to the same DNA sample being present in two different disease cohorts (presumably because the individual suffered with two different diseases) - in such instances both samples were excluded from analysis. Non European ancestry refers to a sample being identified as an outlier from the European population according to principal component analysis. Mixed sample refers to a sample which appeared to be composed of a mixture of two (or more) unique DNA samples. Low signal intensity refers to a sample in which the assay signal intensity was low compared with other samples and with background assay noise. DLRS fail refers to a sample which did not pass the pre-determined QC threshold for the DLRS (derivative log ratio spread) metric. Initial calling quality metric fail refers to a sample that failed one or more of the initial calling QC metrics. Total pre-calling exclusions refers to the total number of samples that were excluded on the basis of information and data available prior to CNV calling. Post-calling quality metric fail refers to a sample that was excluded on the basis of post-calling QC metrics. Duplicates and close relatives refers to a sample that was excluded on the basis of identity or close relatedness with another sample in that cohort according to analysis of similarity of CNV genotype calls. * Note that the CAD pre-calling exclusions include also 405 samples excluded because they were outliers on quality metrics, as described in section 5.1.

Normalisation	PVS scaling	Probe summary	Probes containing SNPs	LDF	CNVs
normalised6	True	PCA	Included	True	524
normalised6	False	PCA	Included	True	236
normalised1	True	Mean	Included	True	289
normalised1	False	Mean	Included	True	227
normalised6	True	Mean	Included	True	197
normalised1	True	PCA	Included	True	461
normalised6	True	PCA	Excluded	True	180
red	True	PCA	Included	True	136
normalised6	True	PCA	Included	False	182
normalised6	False	PCA	Included	False	67
normalised1	True	Mean	Included	False	136
normalised1	False	Mean	Included	False	113
normalised6	True	Mean	Included	False	113
normalised1	True	PCA	Included	False	238
normalised6	True	PCA	Excluded	False	114
red	True	PCA	Included	False	126

Table 9: Analysis pipelines used for calling CNVs in Oxford. normalised6 refers to $log_2(\mathrm{QNorm}(R)/\mathrm{QNorm}(G)+0.5)$. normalised1 refers to $log_2(R/G)$. red refers to using the signal from the Red (sample) channel only. PCA refers to using the first principal component of the probes. PVS refers to probe variance scaling. Mean refers to the mean of the probes. LDF refers to the linear discriminant function proposed in ref. 5. CNVs refers to the number of autosomal CNVs that passed quality control filters, and for which this pipeline was chosen for the CNV. The first line of this table is our standard pipeline.

BD	58C UKBS T2D RA CD CAD HT T1D
T2D	58C UKBS RA CD BD BC T1D
RA	58C UKBS T2D BD CAD HT BC
CD	58C UKBS T2D BD CAD HT BC
CAD	58C UKBS RA CD BD BC T1D
HT	58C UKBS RA CD BD BC T1D
BC	58C UKBS T2D RA CD CAD HT T1D
T1D	58C UKBS T2D BD CAD HT BC

Table 10: List of the Expanded Reference cohorts used for each case cohort.

Cohort	WTCCC2 Affymetrix - $R2 > 0.5$	WTCCC2 Affymetrix - $R2 > 0.8$	WTCCC2 Illumina - $R2 > 0.5$	WTCCC2 Illumina - $R2 > 0.8$	WTCCC1 IMPUTED - $R2 > 0.5$	WTCCC1 IMPUTED - $R2 > 0.8$
CD	1	1	1	0	2	1
T1D	0	0	0	0	0	0
T2D	0	0	0	0	0	0

Table 11: The number of published SNP associations in high LD with CNVs from the other WTCCC studies. The number of highly associated SNPs showing $R2 > 0.5$ and $R2 > 0.8$ are shown for each of three sources of SNP genotypes – imputed genotypes from Affymetrix 500k data from the WTCCC1 study and directly genotyped SNPs from the Affymetrix 6.0 and Illumina 1.2M arrays from the WTCCC2 study.

Plex	Assay	SNP_ID	LD	X2nd.PCRP	X1st.PCRP	UEP_SEQ
RA Plex1	CNVR116.1	rs873308	0.55	ACGTTGGATGCCCACAGCTAGAGAAGCTAA	ACGTTGGATGAGAAGAATTCTCAGCATGTG	CATGAGGACTCAATTAAGC
RA Plex1	CNVR1859.1	rs1510702-Proxy for rs12502699	0.99	ACGTTGGATGGGAGGGAAAACTTCGTTAGC	ACGTTGGATGACCCATAGAACTATGCAACC	CACAAACAGACAAAGGC
RA Plex2	CNVR116.1	rs1053361	0.964	ACGTTGGATGAAGGATGACCCTGAGATGGC	ACGTTGGATGTCTGCTGAGAAGTCCAATC	GAGATGGCTGTCACCAC
RA Plex2	CNVR116.1	rs28553519	0.972	ACGTTGGATGCTTGGCATGCATGGTTGTAG	ACGTTGGATGTCCACTTTCTCATCTCCTGC	GAGGAGGATGAAGGCAA
RA Plex2	CNVR116.1	rs582290	0.974	ACGTTGGATGCACAATCAGGGTAACTGTGC	ACGTTGGATGGGCTACGTGTCCTTCATCAG	CAGTAGTGAGCTGGCCCA
RA Plex2	CNVR116.1	rs1053360	0.994	ACGTTGGATGTCTGCTGAGAAGTCCAATC	ACGTTGGATGAAGGATGACCCTGAGATGGC	GTTCAACACCTACTATGCT
RA Plex2	CNVR3041.1 5'	5' Breakpoint assay	N/A	ACGTTGGATGTCTTGCTGCATCTCTGATCC	ACGTTGGATGGTTAACCAGGAAATCATGTG	CTGATCCATTATTGCTAGAC
RA Plex2	CNVR3041.1 3'	3' Breakpoint assay	N/A	ACGTTGGATGAGATAAAGGGCAAACAAAG	ACGTTGGATGGGAACAGTTTATTCTTATGC	AGGGCAAACAAAGTACAAGTC
RA Plex2	GS35222	Gender assay	N/A	ACGTTGGATGTGCTTCTATGGCCGTTATCC	ACGTTGGATGACTCCAAGTATCCCTGGAAG	ATGGCCGTTATCCCTTTTCAAC
RA Plex2	GS35220	Gender assay	N/A	ACGTTGGATGGCAAAATCATGATAGGATG	ACGTTGGATGGCAGAAATAATGCCAGAGGG	ATGATAGGATGAAATAGTAATACA
CD Plex1	W30576-rs17426195	rs17426195	1	ACGTTGGATGGGACAGCACAGGGCTGAATT	ACGTTGGATGTGGAGTTCTGTTTCTATCGG	ggCAGGGCTGAATTGTTGAT

Table 12: Validation assay primer sequences. Primer sequences are given for Sequenom and PCR assays mentioned in the text and used for follow up of CNV association signals.

CNV3041 RA PCR breakpoint determination
3025_25_F CAGAACTGCCGCATCTTTTT
3025_2627_R CCACTGTAATGCTATGGCTCAA

Chr17 T1D Taqman validation assay
Chr17_Fwd TGTTTCTATCGGACAGCACTTCTTT
Chr17_Rev GACAGCACAGGGCTGAATTG
Chr17_Probe_FAM ATCTGCACTCGATCAA
Chr17_Probe_VIC TCTGCACTCAATCAA

CNVR8164.1 BC validation assay
8164.1_Forward CCAGGTACCACCCAAACTTC
8164.1_Reverse CCATATCAGTCTCTCCAGTCCTCTAA
8164.1_Probe_6-FAM TCGAATCACAGGCAGTGTTGCAGGA

CNVID	Disease	Chromosome	StartCoord	EndCoord	P value Combined Controls	P value Extended Reference	log10 BF Combined Controls	log10 BF Extended Reference	Control MAF	Case MAF	Replication assay platform	Tagging SNP assayed	R 2 of tagging SNP with CNV	Replication cases controls incl WTCCC	Replication cases controls excl WTCCC	Replication Analysis	Replication Results incl WTCCC	Replication Results excl WTCCC	Replication reference number
AC_000138.1_44	T1D				2e-31	2.7e-45	31	45	0.246	0.356	SNP_MA	rs9276162	1.00		3883 / 2649	1		7.33e-50	82
AC_000138.1_44	CD				2.3e-06	3.1e-07	3.3	4.2	0.246	0.207	SNP_MA	rs9275772	1.00	6894 / 7977		2	2.00E-04		58
AC_000138.1_44	RA				0.00083	1.1e-05	1.3	2.7	0.246	0.282	TS(Sequenom)	rs5029394	1.00		3398 / 2743	1		0.001123	NA
CNVR116.1	RA	1	25457812	25537782	0.0062	0.0016	2.7	1.9	0.425	0.391	DA(Sequenom)				3297 / 2465	4		0.431	NA
CNVR1859.1	RA	4	28030554	28031189	0.0013	0.00033	1.5	2.0	0.468	0.492	TS(Sequenom)	rs1510702	0.99		3412 / 2751	3	5	0.8147	NA
CNVR1938.1	T2D	4	61681942	61683010	0.0042	9e-04	1.4	2.0	0.0199	0.0295	SNP_MA	rs930668	1.00	4549 / 5579	2625 / 2641	5	0.0363	0.3487	149
CNVR2522.1	T1D	5	86281754	86282878	7.2e-05	0.00018	2.1	1.7	0.0861	0.0675	SNP_MA	rs1493346	1.00		5594 / 5703	6		0.81	157
CNVR2523.1	BC	5	87414712	87417880	0.026	0.00078	0.7	2.3	0.185	0.199	TS(Illumina 660W)	rs7700867	1.00		1916 / 2769	3		0.6894	NA
CNVR2646.1	CD	5	150157836	150161778	1.1e-07	1.4e-05	5.8	4.1	0.0684	0.0954	SNP_MA	rs1000113	0.99	6894 / 7977		2	7.53E-11		58
CNVR2647.1	CD	5	150183562	150203623	1e-07	4.3e-05	6.1	3.8	0.0708	0.0996	SNP_MA	rs1428555	1.00	6894 / 7977		2	3.86E-10		58
CNVR3041.1	RA	6	113808504	113809828	0.0098	0.024	1.6	2.0	0.0162	0.0229	DA(Sequenom)				3303 / 2824	7		0.484	NA
CNVR3107.1	BC	6	152431681	152433972	0.0013	0.00018	1.8	2.4	0.109	0.128	TS(Sequenom)	rs3778089	1.00		2202 / 3327	3		0.087675	NA
CNVR533.1	T1D	1	228199470	228200805	0.00041	0.013	2.2	1.2	0.176	0.144	SNP_MA	rs10779847	1.00		5594 / 5703	6		0.38	157
CNVR5583.1	T2D	12	69,818,942	69,819,932	3.8e-05	2.5e-06	2.8	4.3	0.382	0.357	SNP_MA	rs1798090	1	4549/ 5579		5	3.9E-05		149
CNVR6454.2	CAD	15	74678304	74682279	0.0021	0.002	1.4	2.1	0.49	0.521	SNP_MA	rs2120108	1.00	4023 / 5880		8	0.0028		46,47,48
CNVR7113.6	T1D	17	40930407	40964305	0.0011	0.00092	1.0	1.2	0.23	0.211	TS(Taqman)	rs17426195	1.00	7911 / 9395	6128 / 6675	9	4.60E-06	0.00024	NA
CNVR7113.6	CD	17	40930407	40964305	0.0018	0.0018	1.4	1.6	0.23	0.212	TS(Sequenom)	rs17426195	1.00	4978 / 6069	3230 / 3131	3	8.60E-05	0.011	NA
CNVR7201.1	BC	17	71873457	71876602	0.00094	0.005	1.7	0.7	0.108	0.0862	TS(Illumina 660W)	rs9889656	1.00		1916 / 2769	3		0.6544	NA
CNVR7543.1	BD	19	12,555,939	12,559,475	1e-04	0.0013	2.9	1.6	0.341	0.298	SNP_MA	rs1864081	0.89	4387 / 6209		10	0.008		27
CNVR8164.1	BC	22	37685753	37722446	0.0012	2.3e-05	1.5	3.1	0.0829	0.097	DA(Illumina 660W)			3660 / 5186	1971 / 3149	3	0.0007	0.1204	NA
WTCCC1_CNVR_1	RA	3	174746835	174772293	1.8e-06	0.0016	3.6	1.7	0.0178	0.036	TS(Sequenom)	rs13061519	0.77		3839 / 2638	3		0.83	NA

Table 13: This table reports the replication results for loci identified in an association analysis of the penultimate data freeze. Details of experimental replication assays are given elsewhere in the supplementary material. NA - not available. (This table is referenced from the main text).

The columns of the table are : P-value, Combined Controls: the p value from the frequentist association test combining UKBS and 58C as controls, P-value, Extended Reference: the p value from the frequentist association test considering UKBS, 58C and aetiologially-unrelated cases as controls, log10(BF) - Combined Controls the log10 of the Bayes Factor from the Bayesian association test combining UKBS and 58C as controls, log10(BF) - Extended Reference the log10 of the Bayes Factor from the Bayesian association test considering UKBS, 58C and aetiologially-unrelated cases as controls, Control MAF , The minor allele frequency in controls (UKBS +58C), Case MAF The minor allele frequency in cases, Replication assay (platform): The type of assay used in replication, if appropriate, the experimental platform used. DA: direct CNV assay, TS: tagging SNP, SNP_MA: SNP meta-analysis, Tagging SNP assayed: If a tagging SNP is used in replication the rsid is given, r2 of tagging SNP with CNV LD between SNP assayed in replication and the CNV, Replication cases/controls (incl. WTCCC): Number of case and controls including samples in the WTCCC experiment, Replication cases/controls (excl. WTCCC): Numbers of cases and controls excluding samples in the WTCCC experiment, Replication Analysis The statistical method used to assess association in replication data (see key below), Replication Results (incl. WTCCC): The p value of the replication association test if WTCCC samples included in a combined test, Replication Results (excl. WTCCC): The p value of the replication association test if only non-WTCCC samples are considered, Replication reference: Paper describing the data used in replication.

Replication analysis key : 1. Genotypic test (2df), 2. Summed Z-scores, 3. Trend Test, 4. CNVtools - trend test of combined signal from 3 probes, 5. Weighted z-statistic-based meta-analysis, fixed effects model, 6. Combination of score statistics across cohorts, 7. Trend test on concordant genotypes from two breakpoint assays, 8. adjusted meta-analysis of data from the WTCCC, German MI I and German MI II studies, 9. Trend test stratified by region, 10. Logistic regression that included 2 ancestry principal components.

Disease	CNV	Chr	Start (bp)	Length (kb)	Fitted number of classes	P value Combined Controls	P value Extended Reference	log10 BF Combined Controls	log10 BF Extended Reference	OR Combined Controls	OR Extended Reference	Control MAF	Case MAF
BD	CNVR73.9	1	13,094,201	18.9	3	0.0027	3.0E-04	3.0	3.6	1.38	1.38		
BD	CNVR4553.4	9	140,139,333	14.2	2	7.3e-05	1.8E-04	2.1	1.8	1.33	1.26	0.06	0.04
BD	CNVR8113.1	22	22,355,609	2.1	3	0.00047	1.5E-03	1.9	2.6	1.19	1.17	0.2	0.17
CAD	CNVR765.1	2	41,817,888	0.9	3	0.005	3.5E-04	1.7	3.0	1.15	1.17	0.43	0.47
CAD	CNVR1152.1	2	226,873,606	5.7	3	0.00025	1.2E-04	2.2	2.8	0.85	0.86	0.34	0.38
CD	CNVR164.1	1	45,407,164	10.9	2	9.3e-06	8.7E-06	1.1	1.2	1.46	1.45	0.01	0
T1D	WTCCC1_CNVR_1	3	174,711,490	75.4	2	8.1e-05	3.4E-02	1.6	0.1	1.4	1.15	0.02	0.03
T1D	CNVR2920.2	6	57,726,662	13.3	2	0.00088	1.1E-05	0.6	2.0	0.79	0.66	0.01	0.02

Table 14: These loci represent the QC-passed loci with moderately significant association test statistics that have not been tested for replication. The thresholds used to define these loci are $p < 1 \times 10^{-4}$ or $\log_{10}(BF) > 2.6$ for either the combined control or expanded reference comparison. (This table is referenced from the main text).

* Expanded Reference test not performed due to association signals at HLA in several other diseases.

Columns:
Fitted number of classes: the number of diploid copy-number classes
P-value Combined Controls: the p value from the frequentist association test combining UKBS and 58C as controls
P-value Extended Reference: the p value from the frequentist association test considering UKBS, 58C and aetiologially-unrelated cases as controls
log10(BF)Combined Controls: the $\log_{10}$ of the Bayes Factor from the Bayesian association test combining UKBS and 58C as controls
log10(BF) Extended Reference: the $\log_{10}$ of the Bayes Factor from the Bayesian association test considering UKBS, 58C and aetiologially-unrelated cases as controls
OR Combined Controls: The odds ratio estimated for each additional copy combining UKBS and 58C as controls
OR Extended Reference: The odds ratio estimated for each additional copy considering UKBS, 58C and aetiologially-unrelated cases as controls
Control MAF: The minor allele frequency in controls (UKBS +58C)
Case MAF: The minor allele frequency in cases

Category	# CNVs	single-class	low MAD
>1 CEU Individual	2741	1143 (42%)	$\sim$ 882 (77%)
1 CEU Individual	3117	1912 (61%)	$\sim$ 1095 (57%)
>1 YRI Individual	1402	1025 (73%)	$\sim$ 516 (50%)
1 YRI Individual	2677	2012 (75%)	$\sim$ 996 (50%)
Novel Inserts	292	126 (43%)	$\sim$ 49 (39%)
Other Sources	385	176 (46%)	$\sim$ 86 (49%)
Total	10614	6394 (60%)	$\sim$ 3624 (57%)

Table 15: Table showing proportion of CNVs called as a single class with evidence of underlying polymorphism. We split the CNVs up into 6 categories according to properties of how they were discovered. The first four rows of the table categorize the CNVs according to how many indviduals showed evidence of the CNV in the GSV analysis [103]. The fifth row details CNVs discovered as novel inserts and the sixth row details all other CNVs not in the previous five rows. The first column details the number of CNVs in each category. The second column gives the number of CNVs in each category that were called as having a single class and this number as a percentage of the total in each category in brackets. The third column lists the number of single-class CNVs which had a low value of the MAD (median absolute deviation) statistic and this number as a percentage of all single-class CNVs in each category. (This table is referenced from the on-line methods).

	Median p-value	
Collection	**Exonic**	**Tagging**
T1D	0.37	0.41
T2D	0.59	0.10
CD	0.45	0.18
CAD	0.52	0.40
BC	0.40	0.66
BD	0.59	0.10
HT	0.64	0.34
RA	0.24	0.18
Control	0.56	0.45

Table 16: P-values for comparisons of groups of CNVs. A non-parametric test was performed comparing the BF distributions between two matched groups of CNVs. The table shows the median p-value across multiple matchings. The BFs are from the comparison of a disease collection against the control collections, except for 'Control' which is from the comparison between the two control collections. Two analyses are shown: *Exonic* compares deletions in exons with deletions not in exons; *Tagging* compares CNVs that are well-tagged by SNPs with those that are poorly tagged. See Section 9.5 for more details.

References

1. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., and Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**(1), 75–81 (2006).

2. Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. Global variation in copy number in the human genome. *Nature* **444**(7118), 444–54 (2006).

3. Fiegler, H., Redon, R., and Carter, N. P. Construction and use of spotted large-insert clone DNA microarrays for the detection of genomic copy number changes. *Nat. Protocols* **2**(3), 577–587 (2007).

4. McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., and Altshuler, D. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).

5. Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D., and Hurles, M. E. A robust statistical method for case-control association testing with copy number variation. *Nat Genet* **40**(10), 1245–52 (2008).

6. Turnbull, C. and Rahman, N. Genetic predisposition to breast cancer: past, present, and future. *Annu Rev Genomics Hum Genet* **9**, 321–45 (2008).

7. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* **358**(9291), 1389–99 (2001).

8. Peto, J. and Mack, T. M. High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet* **26**(4), 411–4 (2000).

9. Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., Ding, W., Bell, R., Rosenthal, J., Hussey, C., Tran, T., McClure, M., Frye, C., Hattier, T., Phelps, R., Haugen-Strano, A., Katcher, H., Yakumo, K., Gholami, Z., Shaffer, D., Stone, S., Bayer, S., Wray, C., Bogden, R., Dayananth, P., Ward, J., Tonin, P., Narod, S., Bristow, P. K., Norris, F. H., Helvering, L., Morrison, P., Rosteck, P., Lai, M., Barrett, J. C., Lewis, C., Neuhausen, S., Cannon-Albright, L., Goldgar, D., Wiseman, R., Kamb, A., and Skolnick, M. H. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**(5182), 66–71 (1994).

10. Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., and Micklem, G. Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**(6559), 789–92 (1995).

11. Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., van Veghel-Plandsoen, M., Elstrodt, F., van Duijn, C., Bartels, C., Meijers, C., Schutte, M., McGuffog, L., Thompson, D., Easton, D., Sodha, N., Seal, S., Barfoot, R., Mangion, J., Chang-Claude, J., Eccles, D., Eeles, R., Evans, D. G., Houlston, R., Murday, V., Narod, S., Peretz, T., Peto, J., Phelan, C., Zhang, H. X., Szabo, C., Devilee, P., Goldgar, D., Futreal, P. A., Nathanson, K. L., Weber, B., Rahman, N., Stratton, M. R., and CHEK2-Breast Cancer Consortium. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* **31**(1), 55–9 (2002).

12. Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., North, B., Jayatilake, H., Barfoot, R., Spanova, K., McGuffog, L., Evans, D. G., Eccles, D., Breast Cancer Susceptibility Collaboration (UK), Easton, D. F., Stratton, M. R.,

and Rahman, N. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* **38**(8), 873–5 (2006).

13. Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Chagtai, T., Jayatilake, H., Ahmed, M., Spanova, K., North, B., McGuffog, L., Evans, D. G., Eccles, D., Breast Cancer Susceptibility Collaboration (UK), Easton, D. F., Stratton, M. R., and Rahman, N. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet* **38**(11), 1239–41 (2006).

14. Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., Jayatilake, H., McGuffog, L., Hanks, S., Evans, D. G., Eccles, D., Breast Cancer Susceptibility Collaboration (UK), Easton, D. F., and Stratton, M. R. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* **39**(2), 165–7 (2007).

15. Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., Wareham, N., Ahmed, S., Healey, C. S., Bowman, R., SEARCH collaborators, Meyer, K. B., Haiman, C. A., Kolonel, L. K., Henderson, B. E., Le Marchand, L., Brennan, P., Sangrajrang, S., Gaborieau, V., Odefrey, F., Shen, C. Y., Wu, P. E., Wang, H. C., Eccles, D., Evans, D. G., Peto, J., Fletcher, O., Johnson, N., Seal, S., Stratton, M. R., Rahman, N., Chenevix-Trench, G., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Garcia-Closas, M., Brinton, L., Chanock, S., Lissowska, J., Peplonska, B., Nevanlinna, H., Fagerholm, R., Eerola, H., Kang, D., Yoo, K. Y., Noh, D. Y., Ahn, S. H., Hunter, D. J., Hankinson, S. E., Cox, D. G., Hall, P., Wedren, S., Liu, J., Low, Y. L., Bogdanova, N., Schurmann, P., Dork, T., Tollenaar, R. A., Jacobi, C. E., Devilee, P., Klijn, J. G., Sigurdson, A. J., Doody, M. M., Alexander, B. H., Zhang, J., Cox, A., Brock, I. W., MacPherson, G., Reed, M. W., Couch, F. J., Goode, E. L., Olson, J. E., Meijers-Heijboer, H., van den Ouweland, A., Uitterlinden, A., Rivadeneira, F., Milne, R. L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Hopper, J. L., McCredie, M., Southey, M., Giles, G. G., Schroen, C., Justenhoven, C., Brauch, H., Hamann, U., Ko, Y. D., Spurdle, A. B., Beesley, J., Chen, X., kConFab, AOCS Management Group, Mannermaa, A., Kosma, V. M., Kataja, V., Hartikainen, J., Day, N. E., Cox, D. R., and Ponder, B. A. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**(7148), 1087–93 (2007).

16. Stacey, S. N., Manolescu, A., Sulem, P., Thorlacius, S., Gudjonsson, S. A., Jonsson, G. F., Jakobsdottir, M., Bergthorsson, J. T., Gudmundsson, J., Aben, K. K., Strobbe, L. J., Swinkels, D. W., van Engelenburg, K. C., Henderson, B. E., Kolonel, L. N., Le Marchand, L., Millastre, E., Andres, R., Saez, B., Lambea, J., Godino, J., Polo, E., Tres, A., Picelli, S., Rantala, J., Margolin, S., Jonsson, T., Sigurdsson, H., Jonsdottir, T., Hrafnkelsson, J., Johannsson, J., Sveinsson, T., Myrdal, G., Grimsson, H. N., Sveinsdottir, S. G., Alexiusdottir, K., Saemundsdottir, J., Sigurdsson, A., Kostic, J., Gudmundsson, L., Kristjansson, K., Masson, G., Fackenthal, J. D., Adebamowo, C., Ogundiran, T., Olopade, O. I., Haiman, C. A., Lindblom, A., Mayordomo, J. I., Kiemeney, L. A., Gulcher, J. R., Rafnar, T., Thorsteinsdottir, U., Johannsson, O. T., Kong, A., and Stefansson, K. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **40**(6), 703–6 (2008).

17. Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W. C., Colditz, G. A., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Hayes, R. B., Tucker, M., Gerhard, D. S., Fraumeni, J. F., J., Hoover, R. N., Thomas, G., and Chanock, S. J. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**(7), 870–4 (2007).

18. Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S. A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A., Aben, K. K., Strobbe, L. J., Albers-Akkers, M. T., Swinkels, D. W., Henderson, B. E., Kolonel, L. N., Le Marchand, L., Millastre, E., Andres, R., Godino, J., Garcia-Prats, M. D., Polo, E., Tres, A., Mouy, M., Saemundsdottir, J., Backman, V. M., Gudmundsson, L., Kristjansson, K., Bergthorsson, J. T., Kostic, J., Frigge, M. L., Geller, F., Gudbjartsson, D., Sigurdsson, H., Jonsdottir, T., Hrafnkels-

son, J., Johannsson, J., Sveinsson, T., Myrdal, G., Grimsson, H. N., Jonsson, T., von Holst, S., Werelius, B., Margolin, S., Lindblom, A., Mayordomo, J. I., Haiman, C. A., Kiemeney, L. A., Johannsson, O. T., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., and Stefansson, K. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **39**(7), 865–9 (2007).

19. Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C. S., Humphreys, M. K., Platte, R., Morrison, J., Maranian, M., Pooley, K. A., Luben, R., Eccles, D., Evans, D. G., Fletcher, O., Johnson, N., dos Santos Silva, I., Peto, J., Stratton, M. R., Rahman, N., Jacobs, K., Prentice, R., Anderson, G. L., Rajkovic, A., Curb, J. D., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Diver, W. R., Bojesen, S., Nordestgaard, B. G., Flyger, H., Dörk, T., Schürmann, P., Hillemanns, P., Karstens, J. H., Bogdanova, N. V., Antonenkova, N. N., Zalutsky, I. V., Bermisheva, M., Fedorova, S., Khusnutdinova, E., SEARCH, Kang, D., Yoo, K. Y., Noh, D. Y., Ahn, S. H., Devilee, P., van Asperen, C. J., Tollenaar, R. A., Seynaeve, C., Garcia-Closas, M., Lissowska, J., Brinton, L., Peplonska, B., Nevanlinna, H., Heikkinen, T., Aittomäki, K., Blomqvist, C., Hopper, J. L., Southey, M. C., Smith, L., Spurdle, A. B., Schmidt, M. K., Broeks, A., van Hien, R. R., Cornelissen, S., Milne, R. L., Ribas, G., González-Neira, A., Benitez, J., Schmutzler, R. K., Burwinkel, B., Bartram, C. R., Meindl, A., Brauch, H., Justenhoven, C., Hamann, U., GENICA Consortium, Chang-Claude, J., Hein, R., Wang-Gohrke, S., Lindblom, A., Margolin, S., Mannermaa, A., Kosma, V. M., Kataja, V., Olson, J. E., Wang, X., Fredericksen, Z., Giles, G. G., Severi, G., Baglietto, L., English, D. R., Hankinson, S. E., Cox, D. G., Kraft, P., Vatten, L. J., Hveem, K., Kumle, M., Sigurdson, A., Doody, M., Bhatti, P., Alexander, B. H., Hooning, M. J., van den Ouweland, A. M., Oldenburg, R. A., Schutte, M., Hall, P., Czene, K., Liu, J., Li, Y., Cox, A., Elliott, G., Brock, I., Reed, M. W., Shen, C. Y., Yu, J. C., Hsu, G. C., Chen, S. T., Anton-Culver, H., Ziogas, A., Andrulis, I. L., Knight, J. A., kConFab, Australian Ovarian Cancer Study Group, Beesley, J., Goode, E. L., Couch, F., Chenevix-Trench, G., Hoover, R. N., Ponder, B. A., Hunter, D. J., Pharoah, P. D., Dunning, A. M., Chanock, S. J., and Easton, D. F. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet* **41**(5), 585–90 (2009).

20. Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G., Hankinson, S. E., Hutchinson, A., Wang, Z., Yu, K., Chatterjee, N., Garcia-Closas, M., Gonzalez-Bosquet, J., Prokunina-Olsson, L., Orr, N., Willett, W. C., Colditz, G. A., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Diver, R., Prentice, R., Jackson, R., Kooperberg, C., Chlebowski, R., Lissowska, J., Peplonska, B., Brinton, L. A., Sigurdson, A., Doody, M., Bhatti, P., Alexander, B. H., Buring, J., Lee, I. M., Vatten, L. J., Hveem, K., Kumle, M., Hayes, R. B., Tucker, M., Gerhard, D. S., Fraumeni, J. F., J., Hoover, R. N., Chanock, S. J., and Hunter, D. J. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* **41**(5), 579–84 (2009).

21. Cox, A., Dunning, A. M., Garcia-Closas, M., Balasubramanian, S., Reed, M. W., Pooley, K. A., Scollen, S., Baynes, C., Ponder, B. A., Chanock, S., Lissowska, J., Brinton, L., Peplonska, B., Southey, M. C., Hopper, J. L., McCredie, M. R., Giles, G. G., Fletcher, O., Johnson, N., dos Santos Silva, I., Gibson, L., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Torres, D., Hamann, U., Justenhoven, C., Brauch, H., Chang-Claude, J., Kropp, S., Risch, A., Wang-Gohrke, S., Schürmann, P., Bogdanova, N., Dörk, T., Fagerholm, R., Aaltonen, K., Blomqvist, C., Nevanlinna, H., Seal, S., Renwick, A., Stratton, M. R., Rahman, N., Sangrajrang, S., Hughes, D., Odefrey, F., Brennan, P., Spurdle, A. B., Chenevix-Trench, G., Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer, Beesley, J., Mannermaa, A., Hartikainen, J., Kataja, V., Kosma, V. M., Couch, F. J., Olson, J. E., Goode, E. L., Broeks, A., Schmidt, M. K., Hogervorst, F. B., Van't Veer, L. J., Kang, D., Yoo, K. Y., Noh, D. Y., Ahn, S. H., Wedrén, S., Hall, P., Low, Y. L., Liu, J., Milne, R. L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Sigurdson, A. J., Stredrick, D. L., Alexander, B. H., Struewing, J. P., Pharoah, P. D., Easton, D. F., and Breast Cancer Association Consortium. A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet* **39**(3), 352–8 (2007).

22. Zheng, W., Long, J., Gao, Y. T., Li, C., Zheng, Y.,

Xiang, Y. B., Wen, W., Levy, S., Deming, S. L., Haines, J. L., Gu, K., Fair, A. M., Cai, Q., Lu, W., and Shu, X. O. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* **41**(3), 324–8 (2009).

23. Müller-Oerlinghausen, B., Berghöfer, A., and Bauer, M. Bipolar disorder. *Lancet* **359**(9302), 241–7 (2002).

24. McGuffin, P., Rijsdijk, F., Andrew, M., Sham, P., Katz, R., and Cardno, A. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch Gen Psychiatry* **60**(5), 497–502 (2003).

25. Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., and Hultman, C. M. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**(9659), 234–9 (2009).

26. Craddock, N. and Sklar, P. Genetics of bipolar disorder: successful start to a long journey. *Trends Genet* **25**(2), 99–105 (2009).

27. Ferreira, M. A., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L., Fan, J., Kirov, G., Perlis, R. H., Green, E. K., Smoller, J. W., Grozeva, D., Stone, J., Nikolov, I., Chambert, K., Hamshere, M. L., Nimgaonkar, V. L., Moskvina, V., Thase, M. E., Caesar, S., Sachs, G. S., Franklin, J., Gordon-Smith, K., Ardlie, K. G., Gabriel, S. B., Fraser, C., Blumenstiel, B., Defelice, M., Breen, G., Gill, M., Morris, D. W., Elkin, A., Muir, W. J., McGhee, K. A., Williamson, R., MacIntyre, D. J., MacLean, A. W., St, C. D., Robinson, M., Van Beck, M., Pereira, A. C., Kandaswamy, R., McQuillin, A., Collier, D. A., Bass, N. J., Young, A. H., Lawrence, J., Ferrier, I. N., Anjorin, A., Farmer, A., Curtis, D., Scolnick, E. M., McGuffin, P., Daly, M. J., Corvin, A. P., Holmans, P. A., Blackwood, D. H., Gurling, H. M., Owen, M. J., Purcell, S. M., Sklar, P., Craddock, N., and Wellcome Trust Case Control Consortium. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* **40**(9), 1056–8 (2008).

28. O'Donovan, M. C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., Nikolov, I., Hamshere, M., Carroll, L., Georgieva, L., Dwyer, S., Holmans, P., Marchini, J. L., Spencer, C. C., Howie, B., Leung, H. T., Hartmann, A. M., Moller, H. J., Morris, D. W., Shi, Y., Feng, G., Hoffmann, P., Propping, P., Vasilescu, C., Maier, W., Rietschel, M., Zammit, S., Schumacher, J., Quinn, E. M., Schulze, T. G., Williams, N. M., Giegling, I., Iwata, N., Ikeda, M., Darvasi, A., Shifman, S., He, L., Duan, J., Sanders, A. R., Levinson, D. F., Gejman, P. V., Cichon, S., Nothen, M. M., Gill, M., Corvin, A., Rujescu, D., Kirov, G., Owen, M. J., Buccola, N. G., Mowry, B. J., Freedman, R., Amin, F., Black, D. W., Silverman, J. M., Byerley, W. F., Cloninger, C. R., and Molecular Genetics of Schizophrenia Collaboration. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* **40**(9), 1053–5 (2008).

29. Green, E. K., Grozeva, D., Jones, I., Jones, L., Kirov, G., Caesar, S., Gordon-Smith, K., Fraser, C., Forty, L., Russell, E., Hamshere, M. L., Moskvina, V., Nikolov, I., Farmer, A., McGuffin, P., Wellcome Trust Case Control Consortium, Holmans, P. A., Owen, M. J., O'Donovan, M. C., and Craddock, N. The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Mol Psychiatry* (2009).

30. International Schizophrenia Consortium, Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., and Sklar, P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**(7256), 748–52 (2009).

31. Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, A. S., Kusenda, M., Malhotra, D., Bhandari, A., Stray, S. M., Rippey, C. F., Roccanova, P., Makarov, V., Lakshmi, B., Findling, R. L., Sikich, L., Stromberg, T., Merriman, B., Gogtay, N., Butler, P., Eckstrand, K., Noory, L., Gochman, P., Long, R., Chen, Z., Davis, S., Baker, C., Eichler, E. E., Meltzer, P. S., Nelson, S. F., Singleton, A. B., Lee, M. K., Rapoport, J. L., King, M. C., and Sebat, J. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**(5875), 539–43 (2008).

32. Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T.,

Buizer-Voskamp, J. E., Hansen, T., Jakobsen, K. D., Muglia, P., Francks, C., Matthews, P. M., Gylfason, A., Halldorsson, B. V., Gudbjartsson, D., Thorgeirsson, T. E., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Bjornsson, A., Mattiasdottir, S., Blondal, T., Haraldsson, M., Magnusdottir, B. B., Giegling, I., Möller, H. J., Hartmann, A., Shianna, K. V., Ge, D., Need, A. C., Crombie, C., Fraser, G., Walker, N., Lonnqvist, J., Suvisaari, J., Tuulio-Henriksson, A., Paunio, T., Toulopoulou, T., Bramon, E., Di Forti, M., Murray, R., Ruggeri, M., Vassos, E., Tosato, S., Walshe, M., Li, T., Vasilescu, C., Mühleisen, T. W., Wang, A. G., Ullum, H., Djurovic, S., Melle, I., Olesen, J., Kiemeney, L. A., Franke, B., GROUP, Sabatti, C., Freimer, N. B., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., Andreassen, O. A., Ophoff, R. A., Georgi, A., Rietschel, M., Werge, T., Petursson, H., Goldstein, D. B., Nöthen, M. M., Peltonen, L., Collier, D. A., St Clair, D., and Stefansson, K. Large recurrent microdeletions associated with schizophrenia. *Nature* **455**(7210), 232–6 (2008).

33. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**(7210), 237–41 (2008).

34. Burbach, J. P. and van der Zwaag, B. Contact in the genetics of autism and schizophrenia. *Trends Neurosci* **32**(2), 69–72 (2009).

35. Zhang, D., Cheng, L., Qian, Y., Alliey-Rodriguez, N., Kelsoe, J. R., Greenwood, T., Nievergelt, C., Barrett, T. B., McKinney, R., Schork, N., Smith, E. N., Bloss, C., Nurnberger, J., Edenberg, H. J., Foroud, T., Sheftner, W., Lawson, W. B., Nwulia, E. A., Hipolito, M., Coryell, W., Rice, J., Byerley, W., McMahon, F., Schulze, T. G., Berrettini, W., Potash, J. B., Belmonte, P. L., Zandi, P. P., McInnis, M. G., Zollner, S., Craig, D., Szelinger, S., Koller, D., Christian, S. L., Liu, C., and Gershon, E. S. Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol Psychiatry* **14**(4), 376–80 (2009).

36. Grozeva, D., Kirov, G., Ivanov, D., Jones, I., Jones, L., Green, E., St Clair, D., Young, A., Ferrier, I., Farmer, A., McGuffin, P., Wellcome Trust Case Control Consortium, Holmans, P., Owen, M., O'Donovan, M., and Craddock, N. Rare copy number variants (CNVs): A point of rarity in genetic risk for bipolar disorder and schizophrenia? *Archives of General Psychiatry* .

37. Spitzer, R. L., Endicott, J., and Robins, E. Research diagnostic criteria: rationale and reliability. *Arch Gen Psychiatry* **35**(6), 773–82 (1978).

38. Wing, J. K., Babor, T., Brugha, T., Burke, J., Cooper, J. E., Giel, R., Jablenski, A., Regier, D., and Sartorius, N. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Arch Gen Psychiatry* **47**(6), 589–93 (1990).

39. McGuffin, P., Farmer, A., and Harvey, I. A polydiagnostic application of operational criteria in studies of psychotic illness. development and reliability of the OPCRIT system. *Arch Gen Psychiatry* **48**(8), 764–70 (1991).

40. Craddock, M., Asherson, P., Owen, M. J., Williams, J., McGuffin, P., and Farmer, A. E. Concurrent validity of the OPCRIT diagnostic system. comparison of OPCRIT diagnoses with consensus best-estimate lifetime diagnoses. *Br J Psychiatry* **169**(1), 58–63 (1996).

41. Green, E. K., Raybould, R., Macgregor, S., Gordon-Smith, K., Heron, J., Hyde, S., Grozeva, D., Hamshere, M., Williams, N., Owen, M. J., O'Donovan, M. C., Jones, L., Jones, I., Kirov, G., and Craddock, N. Operation of the schizophrenia susceptibility gene, neuregulin 1, across traditional diagnostic boundaries to increase risk for bipolar disorder. *Arch Gen Psychiatry* **62**(6), 642–8 (2005).

42. Green, E. K., Raybould, R., Macgregor, S., Hyde, S., Young, A. H., O'Donovan, M. C., Owen, M. J., Kirov, G., Jones, L., Jones, I., and Craddock, N. Genetic variation of brain-derived neurotrophic factor (BDNF) in bipolar disorder: case-control study of over 3000 individuals from the UK. *Br J Psychiatry* **188**, 21–5 (2006).

43. Libby, P. and Theroux, P. Pathophysiology of coronary artery disease. *Circulation* **111**(25), 3481–8 (2005).

44. Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., Lisheng, L., and INTERHEART Study Investigators. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART

study): case-control study. *Lancet* **364**(9438), 937–52 (2004).

45. Lusis, A. J., Mar, R., and Pajukanta, P. Genetics of atherosclerosis. *Annu Rev Genomics Hum Genet* **5**, 189–218 (2004).

46. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145), 661–78 (2007).

47. Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R. J., Meitinger, T., Braund, P., Wichmann, H. E., Barrett, J. H., Konig, I. R., Stevens, S. E., Szymczak, S., Tregouet, D. A., Iles, M. M., Pahlke, F., Pollard, H., Lieb, W., Cambien, F., Fischer, M., Ouwehand, W., Blankenberg, S., Balmforth, A. J., Baessler, A., Ball, S. G., Strom, T. M., Braenne, I., Gieger, C., Deloukas, P., Tobin, M. D., Ziegler, A., Thompson, J. R., Schunkert, H., and WTCCC and the Cardiogenics Consortium. Genomewide association analysis of coronary artery disease. *N Engl J Med* **357**(5), 443–53 (2007).

48. Erdmann, J., Grosshennig, A., Braund, P. S., Konig, I. R., Hengstenberg, C., Hall, A. S., Linsel-Nitschke, P., Kathiresan, S., Wright, B., Tregouet, D. A., Cambien, F., Bruse, P., Aherrahrou, Z., Wagner, A. K., Stark, K., Schwartz, S. M., Salomaa, V., Elosua, R., Melander, O., Voight, B. F., O'Donnell, C. J., Peltonen, L., Siscovick, D. S., Altshuler, D., Merlini, P. A., Peyvandi, F., Bernardinelli, L., Ardissino, D., Schillert, A., Blankenberg, S., Zeller, T., Wild, P., Schwarz, D. F., Tiret, L., Perret, C., Schreiber, S., El Mokhtari, N. E., Schafer, A., Marz, W., Renner, W., Bugert, P., Kluter, H., Schrezenmeir, J., Rubin, D., Ball, S. G., Balmforth, A. J., Wichmann, H. E., Meitinger, T., Fischer, M., Meisinger, C., Baumert, J., Peters, A., Ouwehand, W. H., Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group, Myocardial Infarction Genetics Consortium, Wellcome Trust Case Control Consortium, Cardiogenics Consortium, Deloukas, P., Thompson, J. R., Ziegler, A., Samani, N. J., and Schunkert, H. New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet* **41**(3), 280–2 (2009).

49. Tregouet, D. A., Konig, I. R., Erdmann, J., Munteanu, A., Braund, P. S., Hall, A. S., Grosshennig, A., Linsel-Nitschke, P., Perret, C., DeSuremain, M., Meitinger, T., Wright, B. J., Preuss, M., Balmforth, A. J., Ball, S. G., Meisinger, C., Germain, C., Evans, A., Arveiler, D., Luc, G., Ruidavets, J. B., Morrison, C., van der Harst, P., Schreiber, S., Neureuther, K., Schafer, A., Bugert, P., El Mokhtari, N. E., Schrezenmeir, J., Stark, K., Rubin, D., Wichmann, H. E., Hengstenberg, C., Ouwehand, W., Wellcome Trust Case Control Consortium, Cardiogenics Consortium, Ziegler, A., Tiret, L., Thompson, J. R., Cambien, F., Schunkert, H., and Samani, N. J. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat Genet* **41**(3), 283–5 (2009).

50. Myocardial Infarction Genetics Consortium, Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., Anand, S., Engert, J. C., Samani, N. J., Schunkert, H., Erdmann, J., Reilly, M. P., Rader, D. J., Morgan, T., Spertus, J. A., Stoll, M., Girelli, D., McKeown, P. P., Patterson, C. C., Siscovick, D. S., O'Donnell, C. J., Elosua, R., Peltonen, L., Salomaa, V., Schwartz, S. M., Melander, O., Altshuler, D., Ardissino, D., Merlini, P. A., Berzuini, C., Bernardinelli, L., Peyvandi, F., Tubaro, M., Celli, P., Ferrario, M., Fetiveau, R., Marziliano, N., Casari, G., Galli, M., Ribichini, F., Rossi, M., Bernardi, F., Zonzin, P., Piazza, A., Mannucci, P. M., Schwartz, S. M., Siscovick, D. S., Yee, J., Friedlander, Y., Elosua, R., Marrugat, J., Lucas, G., Subirana, I., Sala, J., Ramos, R., Kathiresan, S., Meigs, J. B., Williams, G., Nathan, D. M., MacRae, C. A., O'Donnell, C. J., Salomaa, V., Havulinna, A. S., Peltonen, L., Melander, O., Berglund, G., Voight, B. F., Kathiresan, S., Hirschhorn, J. N., Asselta, R., Duga, S., Spreafico, M., Musunuru, K., Daly, M. J., Purcell, S., Voight, B. F., Purcell, S., Nemesh, J., Korn, J. M., McCarroll, S. A., Schwartz, S. M., Yee, J., Kathiresan, S., Lucas, G., Subirana, I., Elosua, R., Surti, A., Guiducci, C., Gianniny, L., Mirel, D., Parkin, M., Burtt, N., Gabriel, S. B., Samani, N. J., Thompson, J. R., Braund, P. S., Wright, B. J., Balmforth, A. J., Ball, S. G., Hall, A. S., Wellcome Trust Case Control Consortium, Schunkert, H., Erdmann, J., Linsel-Nitschke, P., Lieb, W., Ziegler, A., Konig, I., Hengstenberg, C., Fischer, M., Stark, K., Grosshennig, A., Preuss, M., Wichmann, H. E., Schreiber, S., Schunkert, H., Samani, N. J., Erdmann, J., Ouwehand, W., Hengstenberg, C.,

Deloukas, P., Scholz, M., Cambien, F., Reilly, M. P., Li, M., Chen, Z., Wilensky, R., Matthai, W., Qasim, A., Hakonarson, H. H., Devaney, J., Burnett, M. S., Pichard, A. D., Kent, K. M., Satler, L., Lindsay, J. M., Waksman, R., Epstein, S. E., Rader, D. J., Scheffold, T., Berger, K., Stoll, M., Huge, A., Girelli, D., Martinelli, N., Olivieri, O., Corrocher, R., Morgan, T., Spertus, J. A., McKeown, P., Patterson, C. C., Schunkert, H., Erdmann, E., Linsel-Nitschke, P., Lieb, W., Ziegler, A., Konig, I. R., Hengstenberg, C., Fischer, M., Stark, K., Grosshennig, A., Preuss, M., Wichmann, H. E., Schreiber, S., Holm, H., Thorleifsson, G., Thorsteinsdottir, U., Stefansson, K., Engert, J. C., Do, R., Xie, C., Anand, S., Kathiresan, S., Ardissino, D., Mannucci, P. M., Siscovick, D., O'Donnell, C. J., Samani, N. J., Melander, O., Elosua, R., Peltonen, L., Salomaa, V., Schwartz, S. M., and Altshuler, D. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* **41**(3), 334–41 (2009).

51. Gudbjartsson, D. F., Bjornsdottir, U. S., Halapi, E., Helgadottir, A., Sulem, P., Jonsdottir, G. M., Thorleifsson, G., Helgadottir, H., Steinthorsdottir, V., Stefansson, H., Williams, C., Hui, J., Beilby, J., Warrington, N. M., James, A., Palmer, L. J., Koppelman, G. H., Heinzmann, A., Krueger, M., Boezen, H. M., Wheatley, A., Altmuller, J., Shin, H. D., Uh, S. T., Cheong, H. S., Jonsdottir, B., Gislason, D., Park, C. S., Rasmussen, L. M., Porsbjerg, C., Hansen, J. W., Backer, V., Werge, T., Janson, C., Jonsson, U. B., Ng, M. C., Chan, J., So, W. Y., Ma, R., Shah, S. H., Granger, C. B., Quyyumi, A. A., Levey, A. I., Vaccarino, V., Reilly, M. P., Rader, D. J., Williams, M. J., van Rij, A. M., Jones, G. T., Trabetti, E., Malerba, G., Pignatti, P. F., Boner, A., Pescollderungg, L., Girelli, D., Olivieri, O., Martinelli, N., Ludviksson, B. R., Ludviksdottir, D., Eyjolfsson, G. I., Arnar, D., Thorgeirsson, G., Deichmann, K., Thompson, P. J., Wjst, M., Hall, I. P., Postma, D. S., Gislason, T., Gulcher, J., Kong, A., Jonsdottir, I., Thorsteinsdottir, U., and Stefansson, K. Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat Genet* **41**(3), 342–7 (2009).

52. Coronary Artery Disease Consortium, Samani, N. J., Deloukas, P., Erdmann, J., Hengstenberg, C., Kuulasmaa, K., McGinnis, R., Schunkert, H., Soranzo, N., Thompson, J., Tiret, L., and Ziegler, A. Large scale association analysis of novel genetic loci for coronary artery disease. *Arterioscler Thromb Vasc Biol* **29**(5), 774–80 (2009).

53. Samani, N. J., Burton, P., Mangino, M., Ball, S. G., Balmforth, A. J., Barrett, J., Bishop, T., Hall, A., and BHF Family Heart Study Research Group. A genomewide linkage study of 1,933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) Family Heart Study. *Am J Hum Genet* **77**(6), 1011–20 (2005).

54. Travis, S. P., Stange, E. F., Lemann, M., Oresland, T., Chowers, Y., Forbes, A., D'Haens, G., Kitis, G., Cortot, A., Prantera, C., Marteau, P., Colombel, J. F., Gionchetti, P., Bouhnik, Y., Tiret, E., Kroesen, J., Starlinger, M., Mortensen, N. J., and European Crohn's and Colitis Organisation (ECCO). European evidence based consensus on the diagnosis and management of Crohn's disease: current management. *Gut* **55 Suppl 1**, i16–35 (2006).

55. Tysk, C., Lindberg, E., Jarnerot, G., and Floderus-Myrhed, B. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. a study of heritability and the influence of smoking. *Gut* **29**(7), 990–6 (1988).

56. Gaya, D. R., Russell, R. K., Nimmo, E. R., and Satsangi, J. New genes in inflammatory bowel disease: lessons for complex diseases? *Lancet* **367**(9518), 1271–84 (2006).

57. Mathew, C. G. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat Rev Genet* **9**(1), 9–14 (2008).

58. Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., NIDDK IBD Genetics Consortium, Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J. P., de Vos, M., Vermeire, S., Louis, E., Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T.,

Prescott, N. J., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M., and Daly, M. J. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**(8), 955–62 (2008).

59. Abraham, C. and Cho, J. H. IL-23 and autoimmunity: new insights into the pathogenesis of inflammatory bowel disease. *Annu Rev Med* **60**, 97–110 (2009).

60. Virgin, H. W. and Levine, B. Autophagy genes in immunity. *Nat Immunol* **10**(5), 461–70 (2009).

61. Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., Radlwimmer, B., and Stange, E. F. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet* **79**(3), 439–48 (2006).

62. McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P., Zody, M. C., Hall, J. L., Brant, S. R., Cho, J. H., Duerr, R. H., Silverberg, M. S., Taylor, K. D., Rioux, J. D., Altshuler, D., Daly, M. J., and Xavier, R. J. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**(9), 1107–12 (2008).

63. Lennard-Jones, J. E. Classification of inflammatory bowel disease. *Scand J Gastroenterol Suppl* **170**, 2–6; discussion 16–9 (1989).

64. Battegay, E., Lip, G., and Bakris, G. *Hypertension: Principles and Practice*. Informa Healthcare, (2005).

65. Kobberling, J. and Tillil, H. Empirical risk figures for first degree relatives of non-insulin dependent diabetics. *The genetics of diabetes mellitus* , 201–209 (1982).

66. Hamet, P. and Seda, O. Current status of genome-wide scanning for hypertension. *Curr Opin Cardiol* **22**(4), 292–7 (2007).

67. Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M. D., Bochud, M., Coin, L., Najjar, S. S., Zhao, J. H., Heath, S. C., Eyheramendy, S., Papadakis, K., Voight, B. F., Scott, L. J., Zhang, F., Farrall, M., Tanaka, T., Wallace, C., Chambers, J. C., Khaw, K. T., Nilsson, P., van der Harst, P., Polidoro, S., Grobbee, D. E., Onland-Moret, N. C., Bots, M. L., Wain, L. V., Elliott, K. S., Teumer, A., Luan, J., Lucas, G., Kuusisto, J., Burton, P. R., Hadley, D., McArdle, W. L., Wellcome Trust Case Control Consortium, Brown, M., Dominiczak, A., Newhouse, S. J., Samani, N. J., Webster, J., Zeggini, E., Beckmann, J. S., Bergmann, S., Lim, N., Song, K., Vollenweider, P., Waeber, G., Waterworth, D. M., Yuan, X., Groop, L., Orho-Melander, M., Allione, A., Di Gregorio, A., Guarrera, S., Panico, S., Ricceri, F., Romanazzi, V., Sacerdote, C., Vineis, P., Barroso, I., Sandhu, M. S., Luben, R. N., Crawford, G. J., Jousilahti, P., Perola, M., Boehnke, M., Bonnycastle, L. L., Collins, F. S., Jackson, A. U., Mohlke, K. L., Stringham, H. M., Valle, T. T., Willer, C. J., Bergman, R. N., Morken, M. A., Döring, A., Gieger, C., Illig, T., Meitinger, T., Org, E., Pfeufer, A., Wichmann, H. E., Kathiresan, S., Marrugat, J., O'Donnell, C. J., Schwartz, S. M., Siscovick, D. S., Subirana, I., Freimer, N. B., Hartikainen, A. L., McCarthy, M. I., O'Reilly, P. F., Peltonen, L., Pouta, A., de Jong, P. E., Snieder, H., van Gilst, W. H., Clarke, R., Goel, A., Hamsten, A., Peden, J. F., Seedorf, U., Syvänen, A. C., Tognoni, G., Lakatta, E. G., Sanna, S., Scheet, P., Schlessinger, D., Scuteri, A., Dörr, M., Ernst, F., Felix, S. B., Homuth, G., Lorbeer, R., Reffelmann, T., Rettig, R., Völker, U., Galan, P., Gut, I. G., Hercberg, S., Lathrop, G. M., Zelenika, D., Deloukas, P., Soranzo, N., Williams, F. M., Zhai, G., Salomaa, V., Laakso, M., Elosua, R., Forouhi, N. G., Völzke, H., Uiterwaal, C. S., van der Schouw, Y. T., Numans, M. E., Matullo, G., Navis, G., Berglund, G., Bingham, S. A., Kooner, J. S., Connell, J. M., Bandinelli, S., Ferrucci, L., Watkins, H., Spector, T. D., Tuomilehto, J., Altshuler, D., Strachan, D. P., Laan, M., Meneton, P., Wareham, N. J., Uda, M., Jarvelin, M. R., Mooser, V., Melander, O., Loos, R. J., Elliott, P., Abecasis, G. R., Caulfield, M., and Munroe, P. B. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* (2009).

68. Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., Glazer, N. L., Morrison, A. C., Johnson, A. D., Aspelund, T., Aulchenko, Y., Lumley, T., Kottgen, A., Vasan,

R. S., Rivadeneira, F., Eiriksdottir, G., Guo, X., Arking, D. E., Mitchell, G. F., Mattace-Raso, F. U., Smith, A. V., Taylor, K., Scharpf, R. B., Hwang, S. J., Sijbrands, E. J., Bis, J., Harris, T. B., Ganesh, S. K., O'Donnell, C. J., Hofman, A., Rotter, J. I., Coresh, J., Benjamin, E. J., Uitterlinden, A. G., Heiss, G., Fox, C. S., Witteman, J. C., Boerwinkle, E., Wang, T. J., Gudnason, V., Larson, M. G., Chakravarti, A., Psaty, B. M., and van Duijn, C. M. Genome-wide association study of blood pressure and hypertension. *Nat Genet* (2009).

69. Ji, W., Foo, J. N., O'Roak, B. J., Zhao, H., Larson, M. G., Simon, D. B., Newton-Cheh, C., State, M. W., Levy, D., and Lifton, R. P. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* **40**(5), 592–9 (2008).

70. Newhouse, S., Farrall, M., Wallace, C., Hoti, M., Burke, B., Howard, P., Onipinla, A., Lee, K., Shaw-Hawkins, S., Dobson, R., Brown, M., Samani, N. J., Dominiczak, A. F., Connell, J. M., Lathrop, G. M., Kooner, J., Chambers, J., Elliott, P., Clarke, R., Collins, R., Laan, M., Org, E., Juhanson, P., Veldre, G., Viigimaa, M., Eyheramendy, S., Cappuccio, F. P., Ji, C., Iacone, R., Strazzullo, P., Kumari, M., Marmot, M., Brunner, E., Caulfield, M., and Munroe, P. B. Polymorphisms in the WNK1 gene are associated with blood pressure variation and urinary potassium excretion. *PLoS One* **4**(4), e5003 (2009).

71. Caulfield, M., Munroe, P., Pembroke, J., Samani, N., Dominiczak, A., Brown, M., Benjamin, N., Webster, J., Ratcliffe, P., O'Shea, S., Papp, J., Taylor, E., Dobson, R., Knight, J., Newhouse, S., Hooper, J., Lee, W., Brain, N., Clayton, D., Lathrop, G. M., Farrall, M., Connell, J., and MRC British Genetics of Hypertension Study. Genome-wide mapping of human loci for essential hypertension. *Lancet* **361**(9375), 2118–23 (2003).

72. Worthington, J., Barton, A., and John, S. The epidemiology of rheumatoid arthritis and the use of linkage and association studies to identify disease genes. In *The Hereditary Basis of Rheumatic Diseases*, 9–28. Birkhauser, Basel (2005).

73. Wordsworth, P. and Bell, J. Polygenic susceptibility in rheumatoid arthritis. *Ann Rheum Dis* **50**(6), 343–6 (1991).

74. Barton, A. and Worthington, J. Genetic susceptibility to rheumatoid arthritis: An emerging picture. *Arthritis Rheum* **61**(10), 1441–1446 (2009).

75. McKinney, C., Merriman, M. E., Chapman, P. T., Gow, P. J., Harrison, A. A., Highton, J., Jones, P. B., McLean, L., O'Donnell, J. L., Pokorny, V., Spellerberg, M., Stamp, L. K., Willis, J., Steer, S., and Merriman, T. R. Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Ann Rheum Dis* **67**(3), 409–13 (2008).

76. Hollox, E. J., Huffmeier, U., Zeeuwen, P. L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P. C., Traupe, H., de Jongh, G., den Heijer, M., Reis, A., Armour, J. A., and Schalkwijk, J. Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet* **40**(1), 23–5 (2008).

77. Aitman, T. J., Dong, R., Vyse, T. J., Norsworthy, P. J., Johnson, M. D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A. J., Petretto, E., Hodges, M. D., Bhangal, G., Patel, S. G., Sheehan-Rooney, K., Duda, M., Cook, P. R., Evans, D. J., Domin, J., Flint, J., Boyle, J. J., Pusey, C. D., and Cook, H. T. Copy number polymorphism in FCGR3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**(7078), 851–5 (2006).

78. Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J. M., Gough, S. C., de Smith, A., Blakemore, A. I., Froguel, P., Owen, C. J., Pearce, S. H., Teixeira, L., Guillevin, L., Graham, D. S., Pusey, C. D., Cook, H. T., Vyse, T. J., and Aitman, T. J. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* **39**(6), 721–3 (2007).

79. Arnett, F. C., Edworthy, S. M., Bloch, D. A., McShane, D. J., Fries, J. F., Cooper, N. S., Healey, L. A., Kaplan, S. R., Liang, M. H., Luthra, H. S., Medsger Jr, T. A., Mitchell, D. M., Neustadt, D. H., Pinals, R. S., Schaller, J. G., Sharp, J. T., Wilder, R. L., and Hunder, G. G. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* **31**(3), 315–24 (1988).

80. MacGregor, A. J., Bamber, S., and Silman, A. J. A comparison of the performance of different methods of disease classification for rheumatoid arthri-

tis. results of an analysis from a nationwide twin study. *J Rheumatol* **21**(8), 1420–6 (1994).

81. Devendra, D., Liu, E., and Eisenbarth, G. S. Type 1 diabetes: recent developments. *BMJ* **328**(7442), 750–4 (2004).

82. Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D. J., Stevens, H., Todd, J. A., Walker, N. M., Rich, S. S., and The Type 1 Diabetes Genetics Consortium. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* **41**, 703–707 (2009).

83. Cooper, J. D., Smyth, D. J., Smiles, A. M., Plagnol, V., Walker, N. M., Allen, J. E., Downes, K., Barrett, J. C., Healy, B. C., Mychaleckyj, J. C., Warram, J. H., and Todd, J. A. 3. *Nat Genet* **40**(12), 1399–401 (2008).

84. Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S. F., Payne, F., Lowe, C. E., Szeszko, J. S., Hafler, J. P., Zeitels, L., Yang, J. H., Vella, A., Nutland, S., Stevens, H. E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L. J., Healy, B., Burren, O. S., Lam, A. A., Ovington, N. R., Allen, J., Adlem, E., Leung, H. T., Wallace, C., Howson, J. M., Guja, C., Ionescu-Tirgoviste, C., Genetics of Type 1 Diabetes in Finland, Simmonds, M. J., Heward, J. M., Gough, S. C., Wellcome Trust Case Control Consortium, Dunger, D. B., Wicker, L. S., and Clayton, D. G. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* **39**(7), 857–64 (2007).

85. Smyth, D. J., Cooper, J. D., Bailey, R., Field, S., Burren, O., Smink, L. J., Guja, C., Ionescu-Tirgoviste, C., Widmer, B., Dunger, D. B., Savage, D. A., Walker, N. M., Clayton, D. G., and Todd, J. A. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat Genet* **38**(6), 617–9 (2006).

86. Hyttinen, V., Kaprio, J., Kinnunen, L., Koskenvuo, M., and Tuomilehto, J. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. *Diabetes* **52**(4), 1052–5 (2003).

87. Clayton, D. G. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet* **5**(7), e1000540 (2009).

88. Field, S. F., Howson, J. M., Maier, L. M., Walker, S., Walker, N. M., Smyth, D. J., Armour, J. A., Clayton, D. G., and Todd, J. A. Experimental aspects of copy number variant assays at CCL3L1. *Nat Med* **15**(10), 1115–7 (2009).

89. Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D., and Todd, J. A. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37**(11), 1243–6 (2005).

90. Ehtisham, S., Hattersley, A. T., Dunger, D. B., Barrett, T. G., and British Society for Paediatric Endocrinology and Diabetes Clinical Trials Group. First uk survey of paediatric type 2 diabetes and MODY. *Arch Dis Child* **89**(6), 526–9 (2004).

91. Owen, K. and Hattersley, A. T. Maturity-onset diabetes of the young: from clinical description to molecular genetic characterization. *Best Pract Res Clin Endocrinol Metab* **15**(3), 309–23 (2001).

92. Sagen, J. V., Raeder, H., Hathout, E., Shehadeh, N., Gudmundsson, K., Baevre, H., Abuelo, D., Phornphutkul, C., Molnes, J., Bell, G. I., Gloyn, A. L., Hattersley, A. T., Molven, A., Søvik, O., and Njølstad, P. R. Permanent neonatal diabetes due to mutations in KCNJ11 encoding Kir6.2: patient characteristics and initial response to sulfonylurea therapy. *Diabetes* **53**(10), 2713–8 (2004).

93. Zimmet, P., Alberti, K. G., and Shaw, J. Global and societal implications of the diabetes epidemic. *Nature* **414**(6865), 782–7 (2001).

94. Stumvoll, M., Goldstein, B. J., and van Haeften, T. W. Type 2 diabetes: principles of pathogenesis and therapy. *Lancet* **365**(9467), 1333–46 (2005).

95. Prokopenko, I., McCarthy, M. I., and Lindgren, C. M. Type 2 diabetes: new genes, new understanding. *Trends Genet* **24**(12), 613–21 (2008).

96. Lyssenko, V., Jonsson, A., Almgren, P., Pulizzi, N., Isomaa, B., Tuomi, T., Berglund, G., Altshuler, D., Nilsson, P., and Groop, L. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med* **359**(21), 2220–32 (2008).

97. Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U., Lamina, C., Lettre, G., Lim, N., Lyon, H. N., McCarroll, S. A., Papadakis, K., Qi, L., Randall, J. C., Roccasecca, R. M., Sanna, S., Scheet, P., Weedon, M. N., Wheeler, E., Zhao, J. H., Jacobs, L. C., Prokopenko, I., Soranzo, N., Tanaka, T., Timpson, N. J., Almgren, P., Bennett, A., Bergman, R. N., Bingham, S. A., Bonnycastle, L. L., Brown, M., Burtt, N. P., Chines, P., Coin, L., Collins, F. S., Connell, J. M., Cooper, C., Smith, G. D., Dennison, E. M., Deodhar, P., Elliott, P., Erdos, M. R., Estrada, K., Evans, D. M., Gianniny, L., Gieger, C., Gillson, C. J., Guiducci, C., Hackett, R., Hadley, D., Hall, A. S., Havulinna, A. S., Hebebrand, J., Hofman, A., Isomaa, B., Jacobs, K. B., Johnson, T., Jousilahti, P., Jovanovic, Z., Khaw, K. T., Kraft, P., Kuokkanen, M., Kuusisto, J., Laitinen, J., Lakatta, E. G., Luan, J., Luben, R. N., Mangino, M., McArdle, W. L., Meitinger, T., Mulas, A., Munroe, P. B., Narisu, N., Ness, A. R., Northstone, K., O'Rahilly, S., Purmann, C., Rees, M. G., Ridderstrle, M., Ring, S. M., Rivadeneira, F., Ruokonen, A., Sandhu, M. S., Saramies, J., Scott, L. J., Scuteri, A., Silander, K., Sims, M. A., Song, K., Stephens, J., Stevens, S., Stringham, H. M., Tung, Y. C., Valle, T. T., Van Duijn, C. M., Vimaleswaran, K. S., Vollenweider, P., Waeber, G., Wallace, C., Watanabe, R. M., Waterworth, D. M., Watkins, N., Wellcome Trust Case Control Consortium, Witteman, J. C., Zeggini, E., Zhai, G., Zillikens, M. C., Altshuler, D., Caulfield, M. J., Chanock, S. J., Farooqi, I. S., Ferrucci, L., Guralnik, J. M., Hattersley, A. T., Hu, F. B., Jarvelin, M. R., Laakso, M., Mooser, V., Ong, K. K., Ouwehand, W. H., Salomaa, V., Samani, N. J., Spector, T. D., Tuomi, T., Tuomilehto, J., Uda, M., Uitterlinden, A. G., Wareham, N. J., Deloukas, P., Frayling, T. M., Groop, L. C., Hayes, R. B., Hunter, D. J., Mohlke, K. L., Peltonen, L., Schlessinger, D., Strachan, D. P., Wichmann, H. E., McCarthy, M. I., Boehnke, M., Barroso, I., Abecasis, G. R., Hirschhorn, J. N., and Genetic Investigation of ANthropometric Traits Consortium (GIANT). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* **41**(1), 25–34 (2009).

98. Bellanne-Chantelot, C., Clauin, S., Chauveau, D., Collin, P., Daumont, M., Douillard, C., Dubois-Laforgue, D., Dusselier, L., Gautier, J. F., Jadoul, M., Laloi-Michelin, M., Jacquesson, L., Larger, E., Louis, J., Nicolino, M., Subra, J. F., Wilhem, J. M., Young, J., Velho, G., and Timsit, J. Large genomic rearrangements in the hepatocyte nuclear factor-1beta (TCF2) gene are the most frequent cause of maturity-onset diabetes of the young type 5. *Diabetes* **54**(11), 3126–32 (2005).

99. Wiltshire, S., Hattersley, A. T., Hitman, G. A., Walker, M., Levy, J. C., Sampson, M., O'Rahilly, S., Frayling, T. M., Bell, J. I., Lathrop, G. M., Bennett, A., Dhillon, R., Fletcher, C., Groves, C. J., Jones, E., Prestwich, P., Simecek, N., Rao, P. V., Wishart, M., Bottazzo, G. F., Foxon, R., Howell, S., Smedley, D., Cardon, L. R., Menzel, S., and McCarthy, M. I. A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *Am J Hum Genet* **69**(3), 553–69 (2001).

100. Frayling, T. M., Walker, M., McCarthy, M. I., Evans, J. C., Allen, L. I., Lynn, S., Ayres, S., Millauer, B., Turner, C., Turner, R. C., Sampson, M. J., Hitman, G. A., Ellard, S., and Hattersley, A. T. Parent-offspring trios: a resource to facilitate the identification of type 2 diabetes genes. *Diabetes* **48**(12), 2475–9 (1999).

101. Groves, C. J., Zeggini, E., Minton, J., Frayling, T. M., Weedon, M. N., Rayner, N. W., Hitman, G. A., Walker, M., Wiltshire, S., Hattersley, A. T., and McCarthy, M. I. Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* **55**(9), 2640–4 (2006).

102. Power, C. and Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* **35**(1), 34–41 (2006).

103. Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., MacArthur, D. G.,

MacDonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. Origins and functional impact of copy number variation in the human genome. *Nature* **advance online publication** (2009).

104. McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., and Altshuler, D. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**(10), 1166–74 (2008).

105. Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tuzun, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**(7191), 56–64 (2008).

106. Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y. H., Frazier, M. E., Scherer, S. W., Strausberg, R. L., and Venter, J. C. The diploid genome sequence of an individual human. *PLoS Biol* **5**(10), e254 (2007).

107. Oeth, P., Beaulieu, M., Park, C. S., Kosman, D., del Mistro, G., van den Boom, D., and Jurinke, C. iPLEX™ assay: Increased plexing efficiency and flexibility for MassARRAY system through single base primer extension with mass-modified terminators. Technical report, Sequenom Application Note, (2005).

108. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res* **12**(4), 656–64 (2002).

109. Agilent Technologies. *Agilent Feature Extraction Software Reference Guide*, volume G4460-90006. Version 9.5 edition, (2007).

110. Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2), 185–93 (2003).

111. Krzanowski, W. and Marriott, F. *Multivariate Analysis, Part 1: Distributions, Ordination and Inference (Kendall's Library of Statistics, No 1)*. Edward Arnold, (1994).

112. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**(3), 375–386 (1955).

113. O'Hagan, A. and Forster, J. *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*, volume 2. Arnold, London, 2nd edition, (2004).

114. Clayton, D. Testing for association on the X chromosome. *Biostatistics* **9**(4), 593–600 (2008).

115. Cardin, N., Vukcevic, D., Pearson, R., Donnelly, P., Marchini, J., and Wellcome Trust Case Control. *A Bayesian Hierarchical Clustering algorithm for CNV data*. Version 9.5 edition, (2009).

116. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**(7), 906–13 (2007).

117. Teo, Y. Y., Inouye, M., Small, K. S., Gwilliam, R., Deloukas, P., Kwiatkowski, D. P., and Clark, T. G. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**(20), 2741–6 (2007).

118. Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhart, A. H., Abraham, C., Regueiro, M., Griffiths, A., Dassopoulos, T., Bitton, A., Yang, H., Targan, S., Datta, L. W., Kistner, E. O., Schumm, L. P., Lee,

A. T., Gregersen, P. K., Barmada, M. M., Rotter, J. I., Nicolae, D. L., and Cho, J. H. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**(5804), 1461–3 (2006).

119. Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J. P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C. A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J. F., Sahbatou, M., and Thomas, G. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**(6837), 599–603 (2001).

120. Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R. H., Achkar, J. P., Brant, S. R., Bayless, T. M., Kirschner, B. S., Hanauer, S. B., Nuñez, G., and Cho, J. H. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**(6837), 603–6 (2001).

121. Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., Vermeire, S., Dewit, O., de Vos, M., Dixon, A., Demarche, B., Gut, I., Heath, S., Foglio, M., Liang, L., Laukens, D., Mni, M., Zelenika, D., Van Gossum, A., Rutgeerts, P., Belaiche, J., Lathrop, M., and Georges, M. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* **3**(4), e58 (2007).

122. Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F. M., Briggs, J., Günther, S., Prescott, N. J., Onnie, C. M., Häsler, R., Sipos, B., Folsch, U. R., Lengauer, T., Platzer, M., Mathew, C. G., Krawczak, M., and Schreiber, S. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet* **39**(2), 207–11 (2007).

123. Rioux, J. D., Xavier, R. J., Taylor, K. D., Silverberg, M. S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M. M., Datta, L. W., Shugart, Y. Y., Griffiths, A. M., Targan, S. R., Ippoliti, A. F., Bernard, E. J., Mei, L., Nicolae, D. L., Regueiro, M., Schumm, L. P., Steinhart, A. H., Rotter, J. I., Duerr, R. H., Cho, J. H., Daly, M. J., and Brant, S. R. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* **39**(5), 596–604 (2007).

124. Rioux, J. D., Daly, M. J., Silverberg, M. S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., Kulbokas, E. J., O'Leary, S., Winchester, E., Dewar, K., Green, T., Stone, V., Chow, C., Cohen, A., Langelier, D., Lapointe, G., Gaudet, D., Faith, J., Branco, N., Bull, S. B., McLeod, R. S., Griffiths, A. M., Bitton, A., Greenberg, G. R., Lander, E. S., Siminovitch, K. A., and Hudson, T. J. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* **29**(2), 223–8 (2001).

125. Parkes, M., Barrett, J. C., Prescott, N. J., Tremelling, M., Anderson, C. A., Fisher, S. A., Roberts, R. G., Nimmo, E. R., Cummings, F. R., Soars, D., Drummond, H., Lees, C. W., Khawaja, S. A., Bagnall, R., Burke, D. A., Todhunter, C. E., Ahmad, T., Onnie, C. M., McArdle, W., Strachan, D., Bethel, G., Bryan, C., Lewis, C. M., Deloukas, P., Forbes, A., Sanderson, J., Jewell, D. P., Satsangi, J., Mansfield, J. C., Wellcome Trust Case Control Consortium, Cardon, L., and Mathew, C. G. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* **39**(7), 830–2 (2007).

126. Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., Jewell, D., Cardon, L., Takazoe, M., Tanaka, T., Ichimori, T., Saito, S., Sekine, A., Iida, A., Takahashi, A., Tsunoda, T., Lathrop, M., and Nakamura, Y. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum Mol Genet* **14**(22), 3499–506 (2005).

127. Kugathasan, S., Baldassano, R. N., Bradfield, J. P., Sleiman, P. M., Imielinski, M., Guthery, S. L., Cucchiara, S., Kim, C. E., Frackelton, E. C., Annaiah, K., Glessner, J. T., Santa, E., Willson, T., Eckert, A. W., Bonkowski, E., Shaner, J. L., Smith, R. M., Otieno, F. G., Peterson, N., Abrams, D. J., Chiavacci, R. M., Grundmeier, R., Mamula, P., Tomer, G., Piccoli, D. A., Monos, D. S., Annese, V., Denson, L. A., Grant, S. F., and Hakonarson,

H. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat Genet* **40**(10), 1211–5 (2008).

128. Villani, A. C., Lemire, M., Louis, E., Silverberg, M. S., Collette, C., Fortin, G., Nimmo, E. R., Renaud, Y., Brunet, S., Libioulle, C., Belaiche, J., Bitton, A., Gaudet, D., Cohen, A., Langelier, D., Rioux, J. D., Arnott, I. D., Wild, G. E., Rutgeerts, P., Satsangi, J., Vermeire, S., Hudson, T. J., and Franchimont, D. Genetic variation in the familial mediterranean fever gene (MEFV) and risk for Crohn's disease and ulcerative colitis. *PLoS One* **4**(9), e7154 (2009).

129. Smyth, D., Cooper, J. D., Collins, J. E., Heward, J. M., Franklyn, J. A., Howson, J. M., Vella, A., Nutland, S., Rance, H. E., Maier, L., Barratt, B. J., Guja, C., Ionescu-Tîrgoviste, C., Savage, D. A., Dunger, D. B., Widmer, B., Strachan, D. P., Ring, S. M., Walker, N., Clayton, D. G., Twells, R. C., Gough, S. C., and Todd, J. A. Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes* **53**(11), 3020–3 (2004).

130. Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M., MacMurray, J., Meloni, G. F., Lucarelli, P., Pellecchia, M., Eisenbarth, G. S., Comings, D., and Mustelin, T. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat Genet* **36**(4), 337–8 (2004).

131. Smyth, D. J., Plagnol, V., Walker, N. M., Cooper, J. D., Downes, K., Yang, J. H., Howson, J. M., Stevens, H., McManus, R., Wijmenga, C., Heap, G. A., Dubois, P. C., Clayton, D. G., Hunt, K. A., van Heel, D. A., and Todd, J. A. Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N Engl J Med* **359**(26), 2767–77 (2008).

132. Ueda, H., Howson, J. M., Esposito, L., Heward, J., Snook, H., Chamberlain, G., Rainbow, D. B., Hunter, K. M., Smith, A. N., Di Genova, G., Herr, M. H., Dahlman, I., Payne, F., Smyth, D., Lowe, C., Twells, R. C., Howlett, S., Healy, B., Nutland, S., Rance, H. E., Everett, V., Smink, L. J., Lam, A. C., Cordell, H. J., Walker, N. M., Bordin, C., Hulme, J., Motzo, C., Cucca, F., Hess, J. F., Metzker, M. L., Rogers, J., Gregory, S., Allahabadia, A., Nithiyananthan, R., Tuomilehto-Wolf, E., Tuomilehto, J., Bingley, P., Gillespie, K. M., Undlien, D. E., Ronningen, K. S., Guja, C., Ionescu-Tîrgoviste, C., Savage, D. A., Maxwell, A. P., Carson, D. J., Patterson, C. C., Franklyn, J. A., Clayton, D. G., Peterson, L. B., Wicker, L. S., Todd, J. A., and Gough, S. C. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**(6939), 506–11 (2003).

133. Nisticò, L., Buzzetti, R., Pritchard, L. E., Van der Auwera, B., Giovannini, C., Bosi, E., Larrad, M. T., Rios, M. S., Chow, C. C., Cockram, C. S., Jacobs, K., Mijovic, C., Bain, S. C., Barnett, A. H., Vandewalle, C. L., Schuit, F., Gorus, F. K., Tosi, R., Pozzilli, P., and Todd, J. A. The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. *Hum Mol Genet* **5**(7), 1075–80 (1996).

134. Nejentsev, S., Howson, J. M., Walker, N. M., Szeszko, J., Field, S. F., Stevens, H. E., Reynolds, P., Hardy, M., King, E., Masters, J., Hulme, J., Maier, L. M., Smyth, D., Bailey, R., Cooper, J. D., Ribas, G., Campbell, R. D., Clayton, D. G., Todd, J. A., and Wellcome Trust Case Control Consortium. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature* **450**(7171), 887–92 (2007).

135. Fung, E. Y., Smyth, D. J., Howson, J. M., Cooper, J. D., Walker, N. M., Stevens, H., Wicker, L. S., and Todd, J. A. Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus. *Genes Immun* **10**(2), 188–91 (2009).

136. Lowe, C. E., Cooper, J. D., Brusko, T., Walker, N. M., Smyth, D. J., Bailey, R., Bourget, K., Plagnol, V., Field, S., Atkinson, M., Clayton, D. G., Wicker, L. S., and Todd, J. A. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat Genet* **39**(9), 1074–82 (2007).

137. Maier, L. M., Lowe, C. E., Cooper, J., Downes, K., Anderson, D. E., Severson, C., Clark, P. M., Healy, B., Walker, N., Aubin, C., Oksenberg, J. R., Hauser, S. L., Compston, A., Sawcer, S., International Multiple Sclerosis Genetics Consortium, De Jager, P. L., Wicker, L. S., Todd, J. A., and Hafler,

D. A. IL2RA genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS Genet* **5**(1), e1000322 (2009).

138. Barratt, B. J., Payne, F., Lowe, C. E., Hermann, R., Healy, B. C., Harold, D., Concannon, P., Gharani, N., McCarthy, M. I., Olavesen, M. G., McCormack, R., Guja, C., Ionescu-Tîrgoviste, C., Undlien, D. E., Rønningen, K. S., Gillespie, K. M., Tuomilehto-Wolf, E., Tuomilehto, J., Bennett, S. T., Clayton, D. G., Cordell, H. J., and Todd, J. A. Remapping the insulin gene/IDDM2 locus in type 1 diabetes. *Diabetes* **53**(7), 1884–9 (2004).

139. Bell, G. I., Horita, S., and Karam, J. H. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33**(2), 176–83 (1984).

140. Hakonarson, H., Qu, H. Q., Bradfield, J. P., Marchand, L., Kim, C. E., Glessner, J. T., Grabs, R., Casalunovo, T., Taback, S. P., Frackelton, E. C., Eckert, A. W., Annaiah, K., Lawson, M. L., Otieno, F. G., Santa, E., Shaner, J. L., Smith, R. M., Onyiah, C. C., Skraban, R., Chiavacci, R. M., Robinson, L. J., Stanley, C. A., Kirsch, S. E., Devoto, M., Monos, D. S., Grant, S. F., and Polychronakos, C. A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes* **57**(4), 1143–6 (2008).

141. Hafler, J. P., Maier, L. M., Cooper, J. D., Plagnol, V., Hinks, A., Simmonds, M. J., Stevens, H. E., Walker, N. M., Healy, B., Howson, J. M., Maisuria, M., Duley, S., Coleman, G., Gough, S. C., International Multiple Sclerosis Genetics Consortium (IMSGC), Worthington, J., Kuchroo, V. K., Wicker, L. S., and Todd, J. A. CD226 Gly307Ser association with multiple autoimmune diseases. *Genes Immun* **10**(1), 5–10 (2009).

142. Gloyn, A. L., Weedon, M. N., Owen, K. R., Turner, M. J., Knight, B. A., Hitman, G., Walker, M., Levy, J. C., Sampson, M., Halford, S., McCarthy, M. I., Hattersley, A. T., and Frayling, T. M. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* **52**(2), 568–72 (2003).

143. Altshuler, D., Hirschhorn, J. N., Klannemark, M., Lindgren, C. M., Vohl, M. C., Nemesh, J., Lane, C. R., Schaffner, S. F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T. J., Daly, M., Groop, L., and Lander, E. S. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* **26**(1), 76–80 (2000).

144. Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., Styrkarsdottir, U., Magnusson, K. P., Walters, G. B., Palsdottir, E., Jonsdottir, T., Gudmundsdottir, T., Gylfason, A., Saemundsdottir, J., Wilensky, R. L., Reilly, M. P., Rader, D. J., Bagger, Y., Christiansen, C., Gudnason, V., Sigurdsson, G., Thorsteinsdottir, U., Gulcher, J. R., Kong, A., and Stefansson, K. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* **38**(3), 320–3 (2006).

145. Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C., and Froguel, P. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**(7130), 881–5 (2007).

146. Sandhu, M. S., Weedon, M. N., Fawcett, K. A., Wasson, J., Debenham, S. L., Daly, A., Lango, H., Frayling, T. M., Neumann, R. J., Sherva, R., Blech, I., Pharoah, P. D., Palmer, C. N., Kimber, C., Tavendale, R., Morris, A. D., McCarthy, M. I., Walker, M., Hitman, G., Glaser, B., Permutt, M. A., Hattersley, A. T., Wareham, N. J., and Barroso, I. Common variants in WFS1 confer risk of type 2 diabetes. *Nat Genet* **39**(8), 951–3 (2007).

147. Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A. M., Ness, A. R., Ebrahim, S., Lawlor, D. A., Ring, S. M., Ben-Shlomo, Y., Jarvelin, M. R., Sovio, U., Bennett, A. J., Melzer, D., Ferrucci, L., Loos, R. J., Barroso, I., Wareham, N. J., Karpe, F., Owen, K. R., Cardon, L. R., Walker, M., Hitman, G. A., Palmer, C. N., Doney, A. S., Morris, A. D., Smith, G. D.,

Hattersley, A. T., and McCarthy, M. I. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**(5826), 889–94 (2007).

148. Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J. T., Thorleifsson, G., Manolescu, A., Rafnar, T., Gudbjartsson, D., Agnarsson, B. A., Baker, A., Sigurdsson, A., Benediktsdottir, K. R., Jakobsdottir, M., Blondal, T., Stacey, S. N., Helgason, A., Gunnarsdottir, S., Olafsdottir, A., Kristinsson, K. T., Birgisdottir, B., Ghosh, S., Thorlacius, S., Magnusdottir, D., Stefansdottir, G., Kristjansson, K., Bagger, Y., Wilensky, R. L., Reilly, M. P., Morris, A. D., Kimber, C. H., Adeyemo, A., Chen, Y., Zhou, J., So, W. Y., Tong, P. C., Ng, M. C., Hansen, T., Andersen, G., Borch-Johnsen, K., Jorgensen, T., Tres, A., Fuertes, F., Ruiz-Echarri, M., Asin, L., Saez, B., van Boven, E., Klaver, S., Swinkels, D. W., Aben, K. K., Graif, T., Cashy, J., Suarez, B. K., van Vierssen Trip, O., Frigge, M. L., Ober, C., Hofker, M. H., Wijmenga, C., Christiansen, C., Rader, D. J., Palmer, C. N., Rotimi, C., Chan, J. C., Pedersen, O., Sigurdsson, G., Benediktsson, R., Jonsson, E., Einarsson, G. V., Mayordomo, J. I., Catalona, W. J., Kiemeney, L. A., Barkardottir, R. B., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., and Stefansson, K. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* **39**(8), 977–83 (2007).

149. Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Bostrom, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burtt, N. P., Chen, H., Chines, P. S., Daly, M. J., Deodhar, P., Ding, C. J., Doney, A. S., Duren, W. L., Elliott, K. S., Erdos, M. R., Frayling, T. M., Freathy, R. M., Gianniny, L., Grallert, H., Grarup, N., Groves, C. J., Guiducci, C., Hansen, T., Herder, C., Hitman, G. A., Hughes, T. E., Isomaa, B., Jackson, A. U., Jorgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F. G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C. M., Lyssenko, V., Marvelle, A. F., Meisinger, C., Midthjell, K., Mohlke, K. L., Morken, M. A., Morris, A. D., Narisu, N., Nilsson, P., Owen, K. R., Palmer, C. N., Payne, F., Perry, J. R., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N. W., Rees, M., Roix, J. J., Sandbaek, A., Shields, B., Sjogren, M., Steinthorsdottir, V., Stringham, H. M., Swift, A. J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N. J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R. M., Weedon, M. N., Willer, C. J., Wellcome Trust Case Control Consortium, Illig, T., Hveem, K., Hu, F. B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N. J., Barroso, I., Hattersley, A. T., Collins, F. S., Groop, L., McCarthy, M. I., Boehnke, M., and Altshuler, D. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**(5), 638–45 (2008).

150. Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., Inouye, M., Freathy, R. M., Attwood, A. P., Beckmann, J. S., Berndt, S. I., Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, Jacobs, K. B., Chanock, S. J., Hayes, R. B., Bergmann, S., Bennett, A. J., Bingham, S. A., Bochud, M., Brown, M., Cauchi, S., Connell, J. M., Cooper, C., Smith, G. D., Day, I., Dina, C., De, S., Dermitzakis, E. T., Doney, A. S., Elliott, K. S., Elliott, P., Evans, D. M., Sadaf Farooqi, I., Froguel, P., Ghori, J., Groves, C. J., Gwilliam, R., Hadley, D., Hall, A. S., Hattersley, A. T., Hebebrand, J., Heid, I. M., Kora, Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H. E., Herrera, B., Hinney, A., Hunt, S. E., Jarvelin, M. R., Johnson, T., Jolley, J. D., Karpe, F., Keniry, A., Khaw, K. T., Luben, R. N., Mangino, M., Marchini, J., McArdle, W. L., McGinnis, R., Meyre, D., Munroe, P. B., Morris, A. D., Ness, A. R., Neville, M. J., Nica, A. C., Ong, K. K., O'Rahilly, S., Owen, K. R., Palmer, C. N., Papadakis, K., Potter, S., Pouta, A., Qi, L., Nurses' Health Study, Randall, J. C., Rayner, N. W., Ring, S. M., Sandhu, M. S., Scherag, A., Sims, M. A., Song, K., Soranzo, N., Speliotes, E. K., Diabetes Genetics Initiative, Syddall, H. E., Teichmann, S. A., Timpson, N. J., Tobias, J. H., Uda, M., SardiNIA Study, Vogel, C. I., Wallace, C., Waterworth, D. M., Weedon, M. N., Wellcome Trust Case Control Consortium, Willer, C. J., FUSION, Wraight, Yuan, X., Zeggini, E., Hirschhorn, J. N., Strachan, D. P., Ouwehand, W. H., Caulfield, M. J., Samani, N. J., Frayling, T. M., Vollenweider, P., Waeber, G., Mooser, V., Deloukas, P., McCarthy, M. I., Wareham, N. J., Barroso, I., Jacobs, K. B., Chanock, S. J., Hayes, R. B., Lamina, C., Gieger, C., Illig,

T., Meitinger, T., Wichmann, H. E., Kraft, P., Hankinson, S. E., Hunter, D. J., Hu, F. B., Lyon, H. N., Voight, B. F., Ridderstrale, M., Groop, L., Scheet, P., Sanna, S., Abecasis, G. R., Albai, G., Nagaraja, R., Schlessinger, D., Jackson, A. U., Tuomilehto, J., Collins, F. S., Boehnke, M., and Mohlke, K. L. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* **40**(6), 768–75 (2008).

151. Prokopenko, I., Langenberg, C., Florez, J. C., Saxena, R., Soranzo, N., Thorleifsson, G., Loos, R. J., Manning, A. K., Jackson, A. U., Aulchenko, Y., Potter, S. C., Erdos, M. R., Sanna, S., Hottenga, J. J., Wheeler, E., Kaakinen, M., Lyssenko, V., Chen, W. M., Ahmadi, K., Beckmann, J. S., Bergman, R. N., Bochud, M., Bonnycastle, L. L., Buchanan, T. A., Cao, A., Cervino, A., Coin, L., Collins, F. S., Crisponi, L., de Geus, E. J., Dehghan, A., Deloukas, P., Doney, A. S., Elliott, P., Freimer, N., Gateva, V., Herder, C., Hofman, A., Hughes, T. E., Hunt, S., Illig, T., Inouye, M., Isomaa, B., Johnson, T., Kong, A., Krestyaninova, M., Kuusisto, J., Laakso, M., Lim, N., Lindblad, U., Lindgren, C. M., McCann, O. T., Mohlke, K. L., Morris, A. D., Naitza, S., Orru, M., Palmer, C. N., Pouta, A., Randall, J., Rathmann, W., Saramies, J., Scheet, P., Scott, L. J., Scuteri, A., Sharp, S., Sijbrands, E., Smit, J. H., Song, K., Steinthorsdottir, V., Stringham, H. M., Tuomi, T., Tuomilehto, J., Uitterlinden, A. G., Voight, B. F., Waterworth, D., Wichmann, H. E., Willemsen, G., Witteman, J. C., Yuan, X., Zhao, J. H., Zeggini, E., Schlessinger, D., Sandhu, M., Boomsma, D. I., Uda, M., Spector, T. D., Penninx, B. W., Altshuler, D., Vollenweider, P., Jarvelin, M. R., Lakatta, E., Waeber, G., Fox, C. S., Peltonen, L., Groop, L. C., Mooser, V., Cupples, L. A., Thorsteinsdottir, U., Boehnke, M., Barroso, I., Van Duijn, C., Dupuis, J., Watanabe, R. M., Stefansson, K., McCarthy, M. I., Wareham, N. J., Meigs, J. B., and Abecasis, G. R. Variants in MTNR1B influence fasting glucose levels. *Nat Genet* **41**(1), 77–81 (2009).

152. Holmkvist, J., Banasik, K., Andersen, G., Unoki, H., Jensen, T. S., Pisinger, C., Borch-Johnsen, K., Sandbaek, A., Lauritzen, T., Brunak, S., Maeda, S., Hansen, T., and Pedersen, O. The type 2 diabetes associated minor allele of rs2237895 KCNQ1 associates with reduced insulin release following an oral glucose load. *PLoS One* **4**(6), e5872 (2009).

153. Grarup, N., Andersen, G., Krarup, N. T., Albrechtsen, A., Schmitz, O., Jørgensen, T., Borch-Johnsen, K., Hansen, T., and Pedersen, O. Association testing of novel type 2 diabetes risk alleles in the jazf1, cdc123/camk1d, tspan8, thada, adamts9, and notch2 loci with insulin release, insulin sensitivity, and obesity in a population-based sample of 4,516 glucose-tolerant middle-aged danes. *Diabetes* **57**(9), 2534–40 (2008).

154. Pritchard, J. K. and Przeworski, M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**(1), 1–14 (2001).

155. Chapman, J. M., Cooper, J. D., Todd, J. A., and Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**(1-3), 18–31 (2003).

156. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin* **1**(6), 80–83 (1945).

157. Wallace, C., Smyth, D. J., Maisuria-Armer, M., Walker, N., Todd, J. A., and Clayton, D. G. The imprinted dlk1-meg3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nat Genet (in press)* (2009).